中国计算机学会学术著作丛书

Reduct 理论

韩素青　赵 岷　著

清华大学出版社

中国计算机学会学术著作丛书

Reduct 理论

Reduct Theory

韩素青　赵 岷　著

清华大学出版社
北京

内 容 简 介

本书系统介绍了基于用户需求的 Reduct 理论。主要内容包括 Reduct 理论、Reduct 典型算法、用户需求描述、基于用户需求的 Reduct 理论、Reduct 与特征选择、数据描述的“规则＋例外”模型以及基于边缘区域的例外分析等。其中数据描述的“规则＋例外”模型源自认知科学，不仅与数据挖掘密切相关，而且与用户需求密切相关。

本书适合从事机器学习、数据挖掘、人工智能、信息处理研究和应用的科技人员学习参考。

图书在版编目(CIP)数据

Reduct 理论/韩素青，赵岷著. --北京：清华大学出版社，2010.4
(中国计算机学会学术著作丛书)
ISBN 978-7-302-21957-6

Ⅰ. ①R…　Ⅱ. ①韩… ②赵…　Ⅲ. ①程序设计－理论研究　Ⅳ. ①TP311

中国版本图书馆 CIP 数据核字(2010)第 018608 号

责任编辑：薛　慧
责任校对：刘玉霞
责任印制：王秀菊

出版发行：清华大学出版社　　**地　　址**：北京清华大学学研大厦 A 座
http://www.tup.com.cn　　**邮　　编**：100084
社　　总　　机：010-62770175　　**邮　　购**：010-62786544
投稿与读者服务：010-62776969，c-service@tup.tsinghua.edu.cn
质　量　反　馈：010-62772015，zhiliang@tup.tsinghua.edu.cn
印 装 者：三河市春园印刷有限公司
经　　销：全国新华书店
开　　本：175×245　**印　张**：24　**字　数**：437 千字
版　　次：2010 年 4 月第 1 版　　**印　　次**：2010 年 4 月第1 次印刷
印　　数：1～2000
定　　价：48.00 元

产品编号：034779-01

清华大学出版社计算机学术著作出版基金

评审委员会

中国计算机学会学术著作丛书

名誉主任委员：张效祥
主任委员：唐泽圣
副主任委员：陆汝钤
委　　员：（以姓氏笔画为序）
王　珊　吕　建　李晓明
林惠民　罗军舟　郑纬民
施伯乐　焦金生　谭铁牛

丛书序

第一台电子计算机诞生于20世纪40年代。到目前为止，计算机的发展已远远超出了其创始者的想象。计算机的处理能力越来越强，应用面越来越广，应用领域也从单纯的科学计算渗透到社会生活的方方面面：从工业、国防、医疗、教育、娱乐直至人们的日常生活，计算机的影响可谓无处不在。

计算机之所以能取得上述地位并成为全球最具活力的产业，原因在于其高速的计算能力、庞大的存储能力以及友好、灵活的用户界面。而这些新技术及其应用有赖于研究人员多年不懈的努力。学术研究是应用研究的基础，也是技术发展的动力。

自1992年起，清华大学出版社与广西科学技术出版社为促进我国计算机科学技术与产业的发展，推动计算机科技著作的出版，设立了“计算机学术著作出版基金”，并将资助出版的著作列为中国计算机学会的学术著作丛书。时至今日，本套丛书已出版学术专著近50种，产生了很好的社会影响，有的专著具有很高的学术水平，有的则奠定了一类学术研究的基础。中国计算机学会一直将学术著作的出版作为学会的一项主要工作。本届理事会将秉承这一传统，继续大力支持本套丛书的出版，鼓励科技工作者写出更多的优秀学术著作，多出好书，多出精品，为提高我国的知识创新和技术创新能力，促进计算机科学技术的发展和进步作出更大的贡献。

中国计算机学会

2002年6月14日

Foreword

In real-world problem solving, it is of paramount importance to use just enough information by focusing on what is essential and ignoring what is irrelevant. Strategies for preventing us from drown in a sea of data have been investigated in wide range of fields under the names of feature construction, feature selection, sampling, data reduction, and so on. This monograph on a reduct theory offers us a novel and fresh view on the classical problem. It describes explicitly, clearly and concisely many concepts and notions in a wide range of different fields, including concept formation and learning, machine learning, data analysis and data mining, pattern analysis, cluster analysis, and many more. The theory provides a much deeper understanding of many intuitive notions.

Although it draws some results from rough sets, the reduct theory emerges as a entirely new theory on its own right. Drs. Suqing Han and Min Zhao's monograph contributes to attribute reduction in several unique ways. First, it provides a comprehensive and thorough review of existing studies in data and attribute reduction and integrates them into a unified reduct theory. Second, it introduces the notion of user-oriented data and attribute reduction. User requirements are formally and flexibly described by an ordering relation on the set of attributes. The relationships between the space of ordering relations and the space of reducts are analyzed. This offers a novel research direction. Third, it develops an attribute-value tree that leads to practical algorithms for attribute reduction. The principles of the second attribute are also explored for producing an efficient reduct construction algorithm. The results make the reduct theory applicable to real-world problems. Fourth, it examines the connection of attribute reduction and a classical topic of feature selection. The reduct theory may be viewed as a theory of feature selection in a wider context. Finally, it demonstrates the value of the theory by applying it in concept learning with a "rule-plus-exception" strategy.

The monograph is a milestone in attribute reduction and feature selection. The research described expands the wisdom in the field. It is distinguished throughout by clear and comprehensive discussions and illustrations. In addition, I can clearly see the presence of not only Professor Jue Wang's philosophy and ways of research but also his wits and unique styles of scientific writing. It is my joy to read the monograph and learn at the same time. It is my great pleasure to introduce you to this original work on attribute reduction and feature selection.

Yiyu Yao
University of Regina

前　言

Pawlak 研究 Rough Sets 的原始动机是针对数据集合(决策表),找到一种介于严格统计学和随意经验之间的能够更好地描述知识不确定性的严谨方法。然而,由于基于 Rough Sets 描述知识不确定性的度量 Roughness 只与给定的决策表相关,因此,Roughness 能否真实描述决策表相对应的问题世界,取决于给定的决策表对于问题世界的概括程度。

在以泛化为目标的统计机器学习中,如果样本集是由对问题世界的有限次观测构成的,那么,该样本集同样需要满足一个表示"概括程度"的条件:样本集合与问题世界同分布,这是讨论泛化问题的基础。Rough Sets 中的 Roughness 与统计学的概率定义在形式上有些相似,但是,两者的本质有所不同,这一点,我们在本书的第 1 章中作了说明。这种本质上的不同导致我们无法从 Roughness 进一步推广并且定义类似于统计学中分布这样的概念,这意味着,Rough Sets 不具备从有限观测(给定数据集合)推测问题世界的理论基础。另外,由于 Roughness 来源于给定的数据集合,因此,这个不确定度量取决于且仅仅取决于给定的数据集合。这一点与模糊集等依赖主观经验的不确定描述方法也不相同。这些原因,使得 Roughness 是否是问题世界的真实描述,不得不依赖于给定数据集合对问题世界的概括程度。更进一步,考虑到 Roughness 所依据的基础——等价关系对数据集合的限制——符号集合,那么把从给定决策表获得的 Roughness 作为描述问题世界知识不确定性的度量,其可信度就更存在疑问了。这是本书为什么没有像大多数有关论著那样以 Rough Sets 为题的原因所在。

Rough Sets 中最重要的贡献是 Reduct,这个概念使得描述不同简洁程度的知识成为可能。这里,简洁程度包括两个含义:其一,属性(变量)的稀疏程度;其二,规则的稀疏程度。对一个给定的决策表,Reduct 理论能够指出哪些属性是必要的或不可替代的,哪些属性是不必要的或可以替代的。这在海量数据以不可控速度不断涌现的今天,有着重要的现实作用。另外,知识描述的简洁程度也可以称为知识粒度,因为人对问题世界的认识程度,往往取决于变量的数量。因此,在假定论域即是问题世界的前提下,如果基于某个 Reduct 对问题世界进行描述,那么有两点值得注意:其一,对建模而言,有些变量是无用的或是可替代的,Reduct 理论恰恰可以指出哪些变量有用,哪些变量无用,并且可以很好地刻画相对最小有用集——Reduct。其

二，对于一个给定的用于描述问题世界的 Reduct，由于 Reduct 中的每一个属性都是必要的，因此，当模型所依赖的属性的数量越少、模型越简单时，随着矛盾样本数量的增加，人们对于问题世界细节的了解就越少。这给予我们一个启示：如果想要获得一个能够凸显问题世界本质的模型，可以在给定 Reduct 的基础上，通过删除一个或几个相对不重要的属性来实现；相应地，如果想要全面了解问题世界，模型至少需要建立在这个给定的 Reduct 之上。

如前所说，由于需要满足比较严苛的条件，Roughness 不能作为描述问题世界的基础，但却可以作为描述给定数据集合简洁程度的度量，也就是“知识粒度”的度量。

Reduct 理论的另一个贡献体现在其理论基础——等价关系之上。尽管目前大多数研究使用一种最简单的等价关系——由样本的属性值相等所诱导的等价关系，但是，这种特殊的等价关系并不是 Reduct 理论必须满足的条件。事实上，Reduct 理论对所有等价关系成立，这是一个不可忽视的特性。换句话说，对于任意一个等价关系，与 Reduct 理论有关的所有结论都可以平行地推广到这个以新等价关系为基础的方法当中。需要指出的是，Reduct 理论本身没有提供任何可以从给定决策表，随意定义等价关系的机制，这是 Reduct 理论的一个缺憾，或许是一个致命的缺憾。

本书围绕用户需求，对 Reduct 计算展开讨论。主要内容包括用户需求描述，基于用户需求计算 Reduct，Reduct 理论与特征选择的关系，基于特征选择框架对 Reduct 算法的全面总结，以及“规则＋例外”模型分析。

最早基于用户需求计算 Reduct 的方法，是由王珏研究员和王驹研究员于 2001 年给出的属性序 Reduct 算法。该算法不仅对 Reduct 完备，而且输出结果唯一。由于算法的计算复杂性与样本的个数成平方关系，而与条件属性的个数呈线性关系，因此，属性序 Reduct 算法通常适用于条件属性个数远大于样本个数的决策表，但不适用于样本个数远大于属性个数的决策表。为解决这个问题，2004 年，赵岷给出了属性-值树 Reduct 算法，并且证明，属性-值树 Reduct 算法与属性序 Reduct 算法关于 Reduct 等价，即，对于同一个属性序，两个算法输出相同的 Reduct。因此，与属性序 Reduct 算法一样，属性-值树 Reduct 算法不仅对 Reduct 完备，而且输出结果唯一。但由于算法的计算复杂性与条件属性的个数成平方关系，而与样本的个数呈线性关系，因此，属性-值树 Reduct 算法通常适用于样本个数远大于条件属性个数的决策表。

如果不涉及用户需求的问题，针对不同类型的数据集合，属性序 Reduct 算法和属性-值树 Reduct 算法都不失为有效的 Reduct 算法，但是如果涉及用户需求的问题，由于这两个算法输出的 Reduct 有时与描述用户需求的属

性序不完全一致，因而是有缺陷的。随后的研究表明，“用户需求可以通过一个属性全序来表示”这种简单的假设还不足以涵盖这类研究，因为许多与此相关的问题还无法明确解释，比如，如何判别一个 Reduct 是否满足用户需求？对于一个给定的属性序，如何评定两个不同的 Reduct 哪个更优？对这些问题的思考，直接导致了关于“用户需求描述问题”的讨论。2006 年，Y. Y. Yao 教授针对用户对于属性的偏好、用户对于属性子集的偏好以及 Reduct 与用户偏好的一致程度等问题，给出了“用户需求描述问题”完整、系统的刻画，这种刻画为用户需求描述以及基于用户需求设计有效算法构建了一个一般性的理论框架。

除属性序 Reduct 算法和属性-值树 Reduct 算法之外，基于自由属性的 Reduct 算法是另外一个基于用户需求计算 Reduct 的算法。该算法于 2001 年由赵凯博士给出，旨在直接获得满足用户需求的 Reduct 算法。就输出结果而言，基于自由属性的 Reduct 算法的确比属性序 Reduct 算法或属性-值树 Reduct 算法有了很大的改进，但遗憾的是，该算法依然无法保证输出的 Reduct 一定与用户的需求完全一致，并且计算复杂性高达决策表样本个数的 4 次方。就计算效率而言，这种计算复杂性显然不能被接受。

为了构建能够直接输出与用户需求完全一致的 Reduct 算法，2004 年，基于属性序 Reduct 算法，韩素青对属性序空间与 Reduct 空间的关系进行了深入研究，获得了一系列直接判定两个属性序 Reduct 是否相同的判定方法。其中最重要的一种判定方法——次属性定理，为随后的研究奠定了基础。

在寻求直接输出满足用户需求 Reduct 算法无果的情况下，我们开始思考这样一个问题：是否存在能够直接输出与用户需求完全一致的 Reduct 算法？令人遗憾的是，答案是否定的。2007 年，借助击中集问题，梁洪力博士证明，计算满足用户需求 Reduct 的问题是 NP-hard 问题。从理论上讲，由于只有直接求解和间接逼近两种方式可以获得满足用户需求的最优解，因此，这个结论迫使研究的目标转向基于次属性定理，在 Reduct 空间中搜索最优解的问题。

从特征选择的角度看，Reduct 计算是一种特殊的特征选择方法。而从 Reduct 计算的角度看，许多特征选择算法则是在执行带有某种偏置的 Reduct 计算，只是计算结果不一定是一个 Reduct 而已。因此，从理论上阐明特征选择与 Reduct 之间的关系，以及基于特征选择框架对目前存在的计算 Reduct 的方法进行全面总结就很有必要。关于这部分内容，遗留的一点遗憾是，由于还未能在理论上将 Reduct 理论与目前统计学中的一个重要研究课题——基于正则化描述对特征进行选择——联系在一起，我们无法将这一主题纳入到本书当中。

基于“规则＋例外”的学习研究源于认知科学，是与用户需求密切相关的另一个研究课题。1998 年，周育健首先将“规则＋例外”的思想融入到 Reduct 理论的研究当中，并在处理 UCI 数据时，获得了一些有启发性的结果。随后，我们逐渐将这类研究理论化，并由此产生了基于边缘区域的“规则＋例外”分析方法。本书没有详细描述周育健的方法，但借鉴了她的思想。本书呈现给读者的有关“规则＋例外”的讨论，都是建立在对边缘区域的分析之上。

在面向用户需求的前提下，本书包含了对 Reduct 理论较为全面并且新颖的研究结果，是我们研究小组近十年研究工作的总结。除作者本人的工作之外，本书收录了王珏、王驹、周育健、赵凯、梁洪力、苗夺谦和崔佳等人的研究成果，在此，作者向他们表示衷心感谢。此外，在征得 Y. Y. Yao 教授同意的条件下，本书还收录了 Y. Y. Yao 教授的相关工作。特别需要指出的是，在本书的写作过程中，作者得到了 Y. Y. Yao 教授的很多帮助，他不仅将最新研究成果提供给作者，而且还仔细阅读了全书并给了重要建议，在此作者也向 Y. Y. Yao 教授表示衷心的感谢。

本书得到国家重大基础研究项目(973 计划)“数字内容理解的理论与方法”子课题“机器学习与数据描述(2004CB318103)”的资助，以及作者单位太原师范学院的支持，在此一并表示感谢。

目　录