文章编号:1000 - 582X(2004)02 - 0044 - 03

粗糙集理论及其在数据挖掘中的应用*

印 勇
(重庆大学 通信工程学院,重庆 400030)

摘 要:Rough sets 理论被广泛运用于不确定环境下的信息处理。基于粗糙集理论的数据挖掘技术正日益受到计算机科学家和数学家的重视。笔者介绍了粗糙集理论的发展过程和基本特点,粗糙集理论在数据挖掘中的应用,以及典型的基于粗糙集的数据挖掘系统,并介绍了粗糙集理论的研究方向和研究领域,最后论述了粗糙集理论与其他智能化方法结合起来处理信息的必要性。

关键词:粗糙集;数据挖掘;知识发现
中图分类号:TP182 **文献标识码**:A

粗糙集理论 RS(Rough set)是波兰数学家 Z. Pawlak 在 1982 年提出的一种分析数据的数学理论[1]。该理论在分类的意义下定义了模糊性和不确定性的概念,是一种处理不确定、不相容数据和不精确问题的新型数学工具。

自 Z. Pawlak 提出粗糙集理论后,粗糙集理论就一直是各国科学家、数学家研究的热点。20 世纪 80 年代,许多波兰学者对粗糙集理论及其应用进行了坚持不懈的深入研究,这个时期广大学者主要是对粗糙集理论的数学性质与逻辑系统进行了深入研究。1991 年粗糙集理论提出者 Z. Pawlak 写出了第一本关于粗糙集的专著,次年 R. Sowindki 主编的粗糙集应用及其与相关方法比较研究的论文集的出版,极大地推动了国际上对粗糙集理论与应用的深入研究。1993 年,在加拿大召开了第一届国际粗糙集与知识发现研讨会。在这次大会上,明确地提出粗糙集理论是进行知识发现(数据挖掘)极好的工具。此后的每一年学术界都举行一届粗糙集理论研讨会,每一届大会都鲜明地提出粗糙集的或者是应用或者是理论的深入发展。值得指出的是在 1995 年的大会上,针对粗糙集理论与模糊集合的基本与相互关系展开了激烈的讨论,极大地推动了粗糙集的进展。1996 年在日本召开了亚洲第一次粗糙集理论大会。此后广大中国学者积极投入到粗糙集理论的研究之中。目前在国内粗糙集理论与知识发现是一个研究的热点,也形成了若干专门的研究机构,如中科院自动化研究所、浙江大学智能信息研究所等;其它的还有很多高校自发形成的研究。如清华大学、西安交通大学、重庆大学等。2001 年在重庆召开了国内第一次粗糙集理论国际研讨大会。

利用粗糙集理论来处理数据挖掘有着较传统挖掘工具不具有的优点。粗糙集理论处理数据不需要对数据的了解,即不再需要对数据的先验信息:比如说统计学中的概率分布、Dempster - Shafer 理论中的概率赋值、或者模糊集理论中的隶属度或概率值;基于粗糙集的数学模型更易于被理解,针对一个特定的大型数据库,利用粗糙集理论比其它理论更容易建立数学模型;许多实验表明,对于同一个数据集,在粗糙集理论工具下进行处理,最终得到的所需的信息更简单、更准确、更易于被决策者接受和理解[2-3]。

目前,粗糙集理论已被广泛应用于人工智能、模式识别和智能信息系统等领域。

1 基于粗糙集的数据挖掘系统[4-7]

近年来,粗糙集理论在数据库领域知识发现(数据挖掘)中的应用取得了较大的进展,基于粗糙集理论的方法逐渐成为数据挖掘主流方法之一。基于粗糙集理论的数据挖掘系统一般都由数据预处理、基于粗糙集或其扩展理论的数据约简、决策算法等部分组成。

* 收稿日期:2003 - 09 - 26
基金项目:重庆市应用基础研究项目(6976)
作者简介:印勇(1963 -),男,重庆人,重庆大学副教授,博士,主要从事数据挖掘和智能信息处理方面的研究。

其大概思想是先进行必要的数据预处理，为数据约简做准备，然后求出约简或近似约简，并在此基础上根据值约简等减少属性和个体数目，最终提取规则并将之应用于新对象的分类。

在过去几年中，国内外建立了不少基于粗糙集的数据挖掘系统。其中最有代表性的有：

1）LERS：LERS（learning from Examples based on Rough Set）系统是美国 Kansas 大学开发的基于粗糙集的实例学习系统。该系统是作为一种开发专家系统的工具被应用的，这种类型的专家系统大多数被用于医疗决策。此外还被应用于环境保护、气候研究和医疗研究。

2）ROSE：波兰 Poznan 科技大学基于粗糙集开发了 ROSE（Rough Set data Explorer）系统，用于决策分析。该系统支持信息系统数据任务，支持新对象的分类，这两个系统已经在许多实际领域中得到应用。而且该系统应用在 windows 平台下。

3）KDD－R：是由加拿大的 Regina 大学开发的基于可变精度粗糙集模型和知识发现的决策矩阵的数据分析系统，本系统被用来对医学数据进行分析，以此产生症状与病症之间新的联系。另外它还支持电信工业的市场研究。

4）ROUGH ENOUGH：本系统是由挪威公司开发的数据挖掘工具。该本系统根据信息系统计算得到可辨识矩阵，并利用许多工具进行集合近似，最后得到约简结果。

2　粗糙集理论的研究方向和研究领域[8-14]

2.1　提出各种简化属性的算法

这是粗糙集理论方面的一个重头戏。对一个巨型的数据集来说，最终的属性可能有成千上万。因此研究快速准确的简化属性是一个非常重要的理论基础。主要的有下面几个算法：

1）基本算法

基本算法首先构造区分矩阵。在区分矩阵的基础上得出区分函数。然后应用吸收律对区分函数进行化简，使之成为析取范式。

2）属性的重要性

由 Hu 提出。该算法非常简单和直观。它使用核作为计算约简的出发点来计算一个最好的或者用户指定的最小约简。该算法将属性的重要性作为启发规则。首先按照属性的重要程度，从大到小逐个加入，直至找到一个最小约简为止。

3）遗传算法

一些科学家提出遗传算法非常适合于进行属性约简。

4）复合系统的约简

该算法的思想是将布尔的化简问题转化成集合睿中的边界搜索问题。

5）扩展法则和动态约简

这两种算法较前面的算法更复杂，然而在大数据集情况下，采用这两种算法却是非常好的算法。

2.2　粗糙集的扩展模型

粗糙集理论应用于数据分析时，会遇到噪音、数据缺少、大数据量等一系列经典理论解决不够理想的问题。因此在近几年的研究中，出现了许多粗糙集的扩展模型。其中典型的有可变精度粗糙集模型、相似模型等。

2.3　粗糙集理论的数学理论研究

随着对粗糙集理论研究的不断深入，与其他数学分支的联系也更加紧密。例如，从算子的观点看粗糙集理论，与之关系较紧的有拓扑空间、数理逻辑、模态逻辑、格与布尔代数、算子代数等。从构造性和集合的观点来看，它与概率论、模糊数学、证据理论、图论、信息论等联系较为密切。粗糙集理论研究不但需要以这些理论作为基础，同时也相应地带动这些理论的发展。由于逻辑是计算机推理的基础，基于粗糙集的逻辑的研究也是粗糙集理论研究的比较活跃的一个方向。

目前，纯粹的数学理论与粗糙集理论结合起来进行研究已有文章发表，并不断有新的数学概念出现，如“粗糙逻辑”、“粗糙理想”、“粗糙半群”等。随着粗糙集理论结构与代数结构，拓扑结构、序结构等各种结构的不断整合，必将不断涌现出新的富有生机的数学分支。

2.4　粗糙集理论与其它不确定性方面的关系及研究

Jelonek 等研究了将 RS 理论用于神经训练数据的预处理，主要进行了属性的缩减和属性值域的缩减，上述处理有利于提高学习效率，并且保持了较低的的近似分类误分数差错率。Hu 等人提出了一种将基于属性的归纳概念方法和 RS 结合的方法，首先使用面向属性的概念树对发生进行泛化，然后使用 RS 方法计算缩减并生成最小属性约简；Lingras 和 Davis 研究了粗糙集和遗传算法的集合，提出了一种粗糙遗传算法。

近年来，又有若干学者提出将其它的不确定方法与粗糙集理论相结合，并提出了相应的算法。其中以粗糙集与模糊集理论相结合的研究为最。

模糊集理论提出之初是为了解决控制领域的难题。1965年,Zadeh提出了模糊集,此后不少计算机科学家和逻辑学家都试图通过这一理论解决G. Frege的含糊概念,然而,遗憾的是模糊集是不可计算的,即模糊集理论不能给出数学公式来描述这一含糊概念,故无法计算出它的隶属函数 u 和模糊逻辑中的算子 y。

模糊集理论和粗糙集理论在处理不确定性和不精确性问题方面都推广了经典集合论。虽然它们有一定的相容性和相似性,然而它们的侧重面不同。从知识的"粒度"的描述上来看,模糊集是通过对象关于集合的隶属程度来描述的,而粗糙集是通过一个集合关于某个可利用的知识库的一对上、下近似来描述的;从集合对象间的关系来看,模糊集强调的是集合边界的病态定义上的,即边界的不分明性,而粗糙集强调的是对象间的不可分辨性;从研究的对象来看,模糊集研究的是属于同一类的不同对象间的隶属关系,重在隶属程度,而粗糙集研究的是不同类中的对象组成的集合关系,重在分类。虽然模糊集的隶属函数和粗糙集的隶属函数都反映了概念的模糊性,直观上有一定的相似性,但是模糊集的隶属函数大多是专家凭经验给出的,因此往往带有很强烈的主观意志,而粗糙集的粗糙隶属函数的计算是从被分析的数据中直接获得的,非常客观。正因为如此,将粗糙集理论和模糊集理论进行某些"整合"后去描述知识的不确定性和不精确性,比它们各自描述知识的不确定性和不精确性可望显示出更强的功能。目前,所见的模糊粗糙集模型中已不乏成功的例子。

总的说来,模糊集理论和粗糙集理论不应被看作是完全不同的两种理论,虽然二者有差别,但是在处理不确定性问题上二者融合在了一起,粗糙集在描述模型方面能力较其它方法为弱,而这恰恰是模糊集理论的强处。另一方面,模糊集理论不能对某问题提出自己的数学公式,这不方便于计算机处理,而利用粗糙集理论这个问题却可以得到很轻松、很简单的解决。因此,综合模糊集理论与粗糙集理论的优缺点,更简单、更高效地处理不确定性、不精确性问题。

3 结 语

粗糙集理论是一种处理不确定和不精确问题的新型数学工具,为数据挖掘提供了一条崭新的途径。粗糙集理论在数据挖掘中的应用研究目前正成为信息科学中的一个研究热点,其发展空间广阔。

参考文献:

[1] PAWLAK Z. Rough sets[J]. Inter J of Computer and Information Sciences, 1982, 11(2): 341 - 356.

[2] 曾黄麟. 粗集理论及其应用[M]. 重庆:重庆大学出版社,1998.

[3] 王国胤. rough集理论与知识获取[M]. 西安:西安交通大学出版社,2001.

[4] 胡可云,王志海,徐本柱. 基于rough Set的知识发现系统[J]. 合肥工业大学学报,1998, 21(1): 71 - 74.

[5] GRZYMALA BAUSSE J W. LERS: A System of Knowledge Discovery Based on Rough Sets[A]. Tsumoto S,ed. Proc of 5th Int. Workshop RSFD96[C]. Tokyo: Morgan Kanfmann RSFD. 1996. 443 - 444.

[6] ZIARKO W, SHAN N. KDD - R: A Comprehensive System for Knowledge Discovery in Databases Using Rough Sets [A]. Lin T Y, Wildberger A M, eds. Soft Computing[C]. San Diego: Simulation Councils Inc, 1995. 298 - 301.

[7] SLOW INSKI R, STEFANOWSKI J. "RoughDAS" and "RoughClass" Software Implementations of the Rough Sets Approach. [M]. Intelligent Decision Support: Handbook of Applications and Advances of Rough Sets Theory. Dordrecht: Kluwer Acasem ic Publishers, 1992. 445 - 456.

[8] KRZYSZTOF J CIOS, ANNA TERESINSKA, STEFANIA KONIECANA, et al. A Knowledge Discovery Approach to Diagnosing Myocardial Perfusion[J]. IEEE EMB Mag,2000,19(4): 17 - 25.

[9] MAN LEUNG WONG, WAI LAM, KWONG SAK LEUNG, et al. Discovering Knowledge From Medical Database Using Evolutionary Algorithms [J]. IEEE EMB Mag, 2000, 19 (4):45 - 55.

[10] SHUSAKU TSUMOTO. Automated Discovery of Positive and Negative Knowledge in Clinical Databases[J]. IEEE EMB Mag, 2000, 19(4):56 - 62.

[11] LINGRAS P J, YAO Y Y, et al. Data Mining Using Extentions of the Rough set Model[J]. Journal of the American society for information science,1998,49(5):415 - 422.

[12] BELL D A, GUAN J W. Computational Methods for Rough Classification and Discovery[J], Journal of the American society for information science,1998,49(5):403 - 414.

[13] GAO S G, REES N W. Identification of Dynamic Fuzzy Models[J]. Fuzzy Sets and Systems, 1995, 74 (5): 307 - 320.

[14] SKOWRON A, NGUYEN H S. Quantization of Real Value Attributes: Rough Set and Boolean Reasoning Approach [R]. Bulletin of International Rough Set Society,1996,1: 5 - 16.

(下转第50页)

参考文献：

[1] 卢福禄. 关注地方财政风险[J]. 了望新闻周刊, 2000, 11(49):12 - 15.

[2] 李俊生. 编制财政风险预算 防范财政风险[N]. 中国财经报, 2000,(2):07 - 11.

[3] JENNINGS N R, SYCARA K P, WOOLDRIAGE M J. Agent technology: Foundations, applications and market [M]. Heidelberg: Springer - Verlag, 1997.

[4] WITTIG T. ARCHON: An architecture for multi - agent systems [M]. Chichester England: Ellis Horwood, 1992.

[5] NILS J, NILSSON. Artifical Intelligence: A new Synthesis. 北京机械工业出版社, 2000.

[6] 王怀民. 基于 Agent 的分布计算环境[J]. 计算机学报, 1996, 10(3):17 - 22.

[7] 张东摩, 李红兵. 人工智能研究动态与发展趋势[J]. 计算机科学, 1998, 25(2): 5 - 8.

Research and Design for Agent - Based Distributing Computation Structure of Local Government Financial Risk Budget Edit System

SHI Wei-ren, JIANG Chang-jiang, PEN Shi-qiang, KANG Jing

(College of Automation, Chongqing University, Chongqing 400030, China)

Abstract: Keeping away and eliminating financial risk is one of the main tasks of every rank government. To solve the problem of local financial risk, a method that constructs a Multi-agent system to edit local financial risk budget is presented. Categorizing the owns of government that give rise to local financial risk, scheming four owns agents, scheming cooperative agent and some auxiliary agents to constitute the system are also proposed. Meanwhile, the agent-based distributing computation structure of the system is given, and the basic structure of agent and the cooperation mechanism of those agents are discussed.

Key words: agent; multiagent system; financial risk budget

(编辑 吕赛英)

(上接第 46 页)

Rough Sets Theory and Its Application to Data Mining

YIN Yong

(College of Communication Engineering, Chongqing University, Chongqing 400030, China)

Abstract: Rough sets is widely applied to dealing with the information under uncertainty. The data mining technology based on the rough sets is being increasingly attached importance by mathematicians and computer scientists. This paper introduces the development process and basic characteristics of rough sets, the application of rough sets theory to the data mining, and the typical system of the data mining based on the rough sets, as well as the directions and fields for future researches. Finally, this paper discusses the necessities through combining rough sets theory and intelligent methods to process information.

Key words: rough sets; data mining; KDD

(编辑 吕赛英)