计算机科学 1997Vol.24№.1

Rough 集*)

Zdzislaw Pawlak, Jerzy Grzymala-Busse, Roman Slowinski, 和 Wojciech Ziarko 著
刘 真译(南昌大学计算机系 南昌 330029)

1993—94 年我在美国 Stanford 大学和 San Jose 州立大学作高访学者时，与本文致谢中提到的 T. Y. Lin 教授合作研究 Rough 集理论及其应用，我们的工作曾在 1994 年 11 月美国举行的第三次 Rough 集和软计算国际学术会议上受到 Z. Pawlak 教授的赞扬。这一理论及其应用现已成立了国际性学术研讨会，参加的成员有波兰、加拿大、美国、日本、俄罗斯、乌克兰和印度等。每年或两年举行一次学术会议。国内也开始陆续发表 Rough 集方面的文章。在此，我们将"Rough 集"译出，以飨读者。

——刘 清

Rough 集理论由 Z. Pawlak 在八十年代初提出[11,12]，是一种处理含糊和不精确性问题的新型数学工具。这种方法似乎对于人工智能(AI)和认知科学是十分重要的，尤其在机器学习、知识获取、决策分析、从数据库的知识发现、专家系统、决策支持系统、归纳推理、模式识别等领域更为重要。

Rough 集概念在某种程度上与许多其它为处理含糊和不精确性问题而研制的数学工具有相似之处，特别是和 Dempster-Shafer 证据理论[15]。两者之间主要区别在于 Dempster-Shafer 理论利用信度函数作为主要工具，而 Rough 集理论利用集合：下近似集和上近似集。另一种关系存在于 fuzzy 集理论和 Rough 集论之间[11]。Rough 集理论与 fuzzy 集理论多方面对照，不是和 fuzzy 竞争，而是补充它[1]。总之，Rough 集理论和 fuzzy 集理论对于不完全的知识来说它们是各自独立的方法。此外，有一些关系存在于 Rough 集理论与辨别式分析之间[7]、与 Bolean 推理方法之间[16]、与决策分析之间[14]。

Rough 集理论的主要优势之一是它不需要任何预备的或额外的有关数据信息，比如说统计学中的概率分布，Dempster-Shafer 理论中基本概率赋值，或者 fuzzy 集理论中的隶属度或概率值。

基本概念

在这篇文章中，我们将假定真实世界的信息是以一种信息表(有时称之为决策表)的形式给出的。因此，信息表表示输入数据，这些数据是从任意领域，诸如医药、财务或军事等领域中收集的。这种信息表的例子见表 1。

表 1 信息表

	属性			决策
	头痛	肌肉痛	体温	流感
e1	是	是	正常	否
e2	是	是	高	是
e3	是	是	很高	是
e4	否	是	正常	否
e5	否	否	高	否
e6	否	是	很高	是

表 1 中的行，标记有 e1,e2,e3,e4,e5 和 e6，被称为实例(个体，实体)。这些实例的性质通过对一些变量的赋值体现出来。我们将识别两种变量：属性(有时称之为条件属性)和决策(有时称之为决策属性)。通常单个决策都要求有全部属性。例如，如果信息表描述一家医院，每个实例可能就是病人，属性是症状和检测，而决策是病症。每一位病人都由检测的结果和症状来表征，而且由医生(专家)根据病症的严重程度被分类。如果这种信息表表示一个工业过程，则这些实例可代表在某些特定时刻及时采集的过程中的样品；属性是过程中的参数；而决策是由操作员(专家)采取的行动。

Rough 集理论的一个主要概念是一种不分明关系，通常与一个属性集合联系在一起，举例来说，这个集合是表 1 中的属性头痛和肌肉痛组成的。实例

*)本文由 Z. Pawlak 教授准许译成中文并在我刊发表。

e1 和 e2 都是由该二种属性相同的值表征的；对于 e1 和 e2 的属性头痛值都是"是"，对于 e1 和 e2 的属性肌肉痛值也都是"是"。此外，实例 e3 从 e1 和 e2 的角度来看是不可分明的，e4 和 e6 同样是互相不能区分的。显然，这种不分明关系是一种等价关系。不分明集被称之为基本集。因此，属性头痛和肌肉痛的集合确定了下面的基本集：{e1,e2,e3}，{e4,e6}和{e5}。任意有限多个基本集的并被称之为可定义集。在我们这种情况中，集合{e1,e2,e3,e5}根据属性头痛和肌肉痛是可定义的，因为我们可以通过属性头痛和肌肉痛同时等于"是"或同时等价于"否"的关系来定义这样的集合。

鉴于不分明关系的概念，使得确定多余(不必要)的属性也就非常简单了。如果一个属性集合和它的包含集定义了相同的不分明关系(即，如果按两种关系划分的基本集都是相同的)，那么属于包含集而不属于该属性集的那些属性是多余的。在表 1 的例子中，设属性集为集合〔头痛，体温〕，而它的包含集是所有三个属性的集合，即集合{头痛，肌肉痛，体温}。由集合{头痛，体温}确定的不分明关系的基本集都是单独一个元素的集合，即{e1}，{e2}，{e3}，{e4}，{e5}，{e6}，而由三个属性构成的集合确定的不分明关系的基本集也如此。因此，属性肌肉痛是多余的。另一方面，集合{头痛，体温}不包含任何多余的属性，因为属性集合{头痛}和{体温}的基本集都不是单独一个元素的集合。如此，一个没有多余属性的属性集被称作最小的(或独立的)。属性集合 P 是另一个属性集合 Q 的简化(或变换)，如果 P 是最小的而且由 P 和 Q 定义的不分明关系是相同的(后一个条件是说由 P 和 Q 定义的不分明关系确定的基本集是相同的)。

在我们所举的例子中，集合{头痛，体温}就是原始属性集合{头痛，肌肉痛，体温}的一个简化。基于这种简化，表 2 介绍了一种新的信息表。

表 2　简化信息表

	属　性		决　策
	头　痛	体　温	流　感
e1	是	正常	否
e2	是	高	是
e3	是	很高	是
e4	否	正常	否
e5	否	高	否
e6	否	很高	是

到目前为止，在我们的讨论中还未包含一个决策。类似于属性，我们可以定义与决策相关联的基本集，即那些拥有相同决策值的全体实例集的子集。这种子集将被称之为概念。对于表 1 和表 2，这些概念是{e1,e4,e5}和{e2,e3,e6}。第一个概念对应于所有未患流感的病人的集合，第二个是所有患流感的病人的集合。问题是我们是否要在表 2 的属性值的基础上陈述谁未患流感和谁患流感。为回答这个问题，我们可以用 Rough 集理论方法观察，决策"流感"取决于属性头痛和体温，因为所有与集{头痛，体温}相关联的不分明关系的基本集都是某些概念的子集。事实上，从表 2 可归纳出下面的规则：

(体温，正常)→(流感，否)；
(头痛，否)且(体温，高)→(流感，否)；
(头痛，是)且(体温，高)→(流感，是)；
(体温，很高)→(流感，是)。

表 3　不一致信息表

	属　性		决　策
	头　痛	体　温	流　感
e1	是	正常	否
e2	是	高	是
e3	是	很高	是
e4	否	正常	否
e5	否	高	否
e6	否	很高	是
e7	否	高	是
e8	否	很高	否

现在，在表 2 的数据项基础上增加二个额外的实例 e7 和 e8，如表 3 所示。由属性头痛和体温定义的不分明关系的基本集是{e1}，{e2}，{e3}，{e4}，{e5,e7}，和{e6,e8}，而由决策"流感"定义的概念是集{e1,e4,e5,e8}和{e2,e3,e6,e7}。

很明显，在表 3 中决策"流感"不取决于属性"头痛"和"体温"，因为不论{e5,e7}还是{e6,e8}都不是任何概念的子集。换句话说，两概念的任一个不是属性集{头痛，体温}可定义的。我们说表 3 是不一致的，因为实例 e5 和 e7 是冲突的(或者说不一致)，这两个实例的任一属性值对应相同，然而其决策值却不同(实例 e6 和 e8 也是冲突的)。

在这种情况下，Rough 集理论提供了一种处理不一致性的工具。这个思想很简单，即对每个概念 X，包含于 X 中的最大可定义集和包含 X 的最小可定义集都是能计算的。前者称为 X 的下近似集，后

者称为X的上近似集。在表3的情况下，概念集{e2，e3，e6，e7}描述了患流感的人，其下近似集等于{e2，e3}且上近似集等于{e2，e3，e5，e6，e7，e8}，如图1所示。

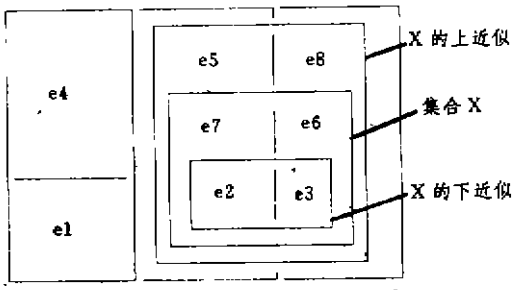

图1 合集X的下近似和上近似

类似地，对于概念{e1，e4，e5，e8}，其下近似集是{e1，e4}且上近似集是{e1，e4，e5，e6，e7，e8}(注：译者根据上下文将原文的e2和e3分别修改为e1和e4)。这两个概念集中任何一个都是Rough集的一个例，也就是，它是一个用给定属性不可定义的集合。包含于X的上近似而非X的下近似中元素的集合{e5，e6，e7，e8}，被称为是边界区域。边界区域中的元素不能被分类成集合X中的成员。另一方面，Rough集也可以被定义为有非空边界区域的集合。

对于任何一个概念，由它的下近似引出的规则必然是有效的(因此，这样的规则称为必然的)，从该概念的上近似引出的规则是可能有效的(称之为可能的)。对表3来说，必然规则是：

(体温，正常)→(流感，否)；
(头痛，是)且(体温，高)→(流感，是)；
(头痛，是)且(体温，很高)→(流感，是)。

可能规则是：

(头痛，否)→(流感，否)；
(体温，正常)→(流感，否)；
(体温，高)→(流感，是)；
(体温，很高)→(流感，是)。

Rough集理论被用来研制一些不精确性的量度。最常用的是，下近似的特性和上近似的特性。对于一给定的实例集合X，不需要说明它是用属性集合P可定义的，其下近似特性是下近似中所有元素的个数与所有实例总数的比值。同样，上近似特性是上近似中元素个数与所有实例总个数的比值。因此，在表3的例中，对于概念集X={e1，e4，e5，e8}，下近似特性是0.25，而上近似特性是0.75。

下近似特性可以被解释为在X中按P中属性所有必然分类的实例数目与信息表中总实例数的比值。这是一种相对的频率。而且，下近似特性按照Dempster-Shafer理论，它是一个信度函数。同样，上近似特性是在X中根据P中属性所有可能分类的实例数目与系统中所有实例总数的比值。因此，它也是一种相对频率。上近似特性按照Dempster-Shafer理论观点，它是一种似然函数[3]。Rough集理论是客观的，即对于一给定的信息表，相应的近似特性是可以被计算的。另一方面，Dempster-Shafer理论是主观的，即信度值(或似然度)被假定是由某个专家给出的。(有关Rough集理论和Dempster-Shafer理论之间的关系的更多信息，参看文[15])。

Rough集合的应用

Rough集理论已经证实在实践中是非常有用的，从许多现实生活中应用的记录来看已经非常明显。一些有关实际操作的描述可参看文[9]、[18]、[24]。然而，Rough集方法学中的大部分应用仅就公开发表的资料还不能完全反映出来，由于冗长的实验数据和新软件的开发，使得不少应用还处于改进之中。

利用Rough集理论处理的主要问题包括数据简化(即删除多余的数据)，数据相关性的发现，数据意义的评估，由数据产生决策(控制)算法，数据的近似分类，数据中的相似或差异的发现，数据中范式的发现，以及因-果关系的发现。

特别地，Rough集方法已经被发现在医学、药学、商业、金融、市场研究、工程设计、气象学、振动分析、开关函数、冲突分析、图像处理、声音识别、并发系统分析、决策分析、字符识别及其它领域都有重要的应用。

Rough集，知识获取和机器学习

在规则的形式中，通过从训练的例子学习而导出的知识，可以被用在基于规则的专家系统中。这些规则比起包含于原始输入数据的信息来说更为一般，因为还未与原始数据的例子匹配的新例子可以通过这些规则正确地被分类。

单凭经验的学习系统被称为基于Rough集例子的学习(LERS)，这是Kansas大学开发的，它包括两种示例学习选项和两种知识获取选项[3,4]。机器学习选项产生一个充分的包括信息表中所有实例的规则集。知识获取选项产生更大的全体规则集。这些规则是通过一张信息表给定的输入数据按给定的选项

导出的。象文[5]中所显示的一样，当一个专家系统必须处理不完全信息时，机器学习的方法作为一个知识获取的工具是不充分的。LERS 系统的知识获取的选项对于用不完全信息工作的专家系统建立知识库是一个恰当的规则归纳法的例子。

LERS 可以从信息表形式中给定的实例导出一套规则集，并且可以利用这一套规则分类新的实例。首先，LERS 检验输入数据的一致性。如果数据不一致，则每个概念集的下和上近似都被计算[4]。现在用户在两种机器学习选项和两种知识获取选项之间提供一种选择。如果机器学习选项被利用，则这系统对每个概念导出一个单一的最小的判别式描述。如果知识获取选项被运用，则导出一个完全的规则集。

系统 LERS 已经为 NASA 的 Johnson 空间中心应用了两年，它是作为一种开发专家系统的工具被引用的，这种类型的专家系统大多数可能被用于空间站释放的板上医疗决策。

LERS 的另一个应用是标题 Ⅲ 的 311，312 和 313 节中介绍的 Emergency Planning 和 Community Right to Know 部门应用的增强设施。这项计划是由美国环境保护机构资助的。

LERS 还被应用于两项医学方面：其一用来比较手术后的病人取暖设备的效果，其二用来评估孕妇的超强度劳动的危险。超产期的预测是一个知识贫乏的领域。现有的超产期的人工估计方法有 17%-38%的准确率。机器学习系统被用于三个不同的关于孕妇的数据集。由 LERS 引入的规则连同 LERS 的分类模式一起共同被应用，并且是在遗传算法并通过部分匹配得到改进的一个“桶队算法”的基础上应用。在新的未见过的情况中超产期的结果预测精确度高得多(68%-90%)。

还有一种很重要的 LERS 用途是全球气候变化的研究。描述对全球气温有影响的规则由一些属性所表征的数据引出，比如太阳的能量释放、火山活动、美国南部的指针摇摆器、二氧化碳流向和二氧化碳的余量。这方面的专家依据获得的新数据把握地球气候变化的奥妙，正如[6]中报告的一样。

Rouah 集和决策分析

波兰 Poznan 科技大学在所开发的称之为 Rough DAS 和 Rough Class 的计算机系统中已经实现了决策分析的 Rough 集方法。它们对任务分别执行解释和描述。这两个系统已经在许多实际领域都有应用[18]。

在医疗方面的另一个应用是用高精度选择迷走神经切断术(HSV)方法处理十二指肠溃疡时所涉及到的指示器指示量度的验证。用 Rough 集在用 11 个关于手术属性描述的 122 个患者的集合上操作，这个描述减少到只用五个相关属性就能保证得到一个能被接受的分类特性。这些减少的属性是基于能显示影响病人的负效应的测试的，从对 70 个新病人的手术的好坏结果类的下近似得到的 44 条决策规则的应用给出了一个 HSV 结果的好的增长值，即从 82%到 93%[17]。

另一个应用涉及到化学结构和 201 四咪唑合成的抗菌剂活动之间关系的分析。这些合成用 8 个涉及结构的属性描述且被分成五个活动类。用 Rough-DAS 系统发现这些属性的简化是由四个属性组成。两个决策规则集，一个有 22 条规则，另一个有 35 条规则，依据相关的结构特性，关于如何设计新的抗菌剂活动合成要给出明晰的建议[8]。

在技术诊断方面，Rough 集理论已被应用于摆动症状的诊断能力的分析。对于确定滚珠轴承诊断中所用的噪声和摆动两者属性的极限值，客观地比较不同的方法，Rough 集似是一种好的工具。作为这项研究的一个结果，已确定摆动症状优于噪声症状，且滚动轴承的技术状态的分类也被建立，它利用 3 条相关症状，14 条决策规则(而不是 12 条)组成[10]。

在经济领域，Rough 集方法已被用于鉴别一个面临破产危机的厂商。为了确定让厂商信赖保险单的金额，分析希腊工业发展银行 ETEVA 的一个实际经验。从这个分析得到的结果用实例支持的明显的规则表达，很受经济专家的欣赏。定性的(无规律的)和定量的属性同时被考虑到了，在传统的多准则决策制定方法中，值函数的构造是一项困难的任务。只利用初始表中 5%-7%的条件就得到决策规则的集合。

现在正在进行的一项单凭经验为依据的研究，它涉及到分类脑部肿瘤微型图片神经网络中的数据简化。已经发现学习时加快了数据简化高达 4.72 倍，这项研究的另一方面的结果是，隐蔽层中神经元的最小数目等于最小减少量的基数。这些有希望的结果显示了 Rough 集方法对神经网络中的数据预处理是一个很有用的工具。

Rough 集和知识发现

象前面所指出的一样，实际上实现了的 Rough 集方法学的应用包括了一系列广阔的区域。引起越

来越多关注的一项重要应用是数据库中的知识发现(KDD)。知识发现或数据库的挖掘是人工智能的一个相对新的子领域，它涉及到从不断增长的企业信息数据库中挖掘出额外的非平凡的知识方面。在这个内容上，主要任务之一是内部数据的关联或关系的发现和特征；例如，医学数据库中症状和病症之间，出现在这种关系中的基本因素的揭示描述将帮助用户更好地理解关于被收集数据的现象的本质，或者说它能被用来进行预测[20,21,24,25]。另一方面，运用 Rough 集方法，还能处理的另一方面是对数据中的反常模式或行为的发现，目的是为了检测假的或非法进入的数据[23]。

KDD 的方法学主要源于早先统计学、数据库理论和机器学习领域中的研究[22]。然而，Rough 集领域中的更新发展确立了这一方法学成为 KDD 问题的主要方法之一[23-25]。

Rough 集技术被用于相关 KDD 研究目的已经五年了。特别是，对数据库的挖掘，比如说 Datalogic[20]，基于 PC 的商业软件系统的有效性使得这种技术从工业和科学的不同部门的用户都是有益的。当前，Rough 集方法学正在其它领域中被引用：市场研究[25]、医学数据分析[17]、药物研究[8]、以控制为目的的传感器数据分析、以及导致新的合成材料设计的研究。库存市场数据的分析已证明了一些著名的市场规则，并导致了一些新的重要规则的发现[25]。利用 Rough 集原始模型扩充的知识发现方法学称为可变精度 Rough 集，Regian 大学在最新的基于工作站的工具集中已经实现了关于知识发现的决策矩阵方法[23]，称其为 KDD-R，KDD-R 被用来对医学数据分析并且当前正支持电信工业的市场研究。

结论 Rough 集方法学已经证明了它在许多实际生活应用中是完备的和十分有用的。Rough 集理论提供了在 AI 的许多分枝上可应用的有效方法。Rough 集理论的有利条件之一是实现它的方法的程序可以很容易地在并行计算机上运行。

然而，许多问题仍尚待解决。尽管 Rough 集理论产生于真正的数学基础，但许多理论问题仍有待于真正澄清。Rough 逻辑，一种基于 Rough 集哲学的不精确推理的逻辑，它似乎是最重要的课题。基于 Rough 集理论对神经网络和遗传算法方法的开发也似乎是很重要的。Rough 控制，即基于 Rough 集理论的控制也似乎是一个非常有前途的应用领域。然而，基于 Rough 集哲学的一个定性的控制理论必须被建立。Rough 集理论与非标准分析、非参数统计学及定性物理学之间的关系是另一些重要课题。

致谢 作者在此对 T. Y. Lin 的帮助和鼓励表示谢意。

译自《Comm. of the ACM》,38(11)1995,PP94-95)

(上接第 37 页)

[6]Thekkath & H. M. Levy, Limits to Low-Latency Communication on High-Speed Networks, ACM Trans. Comput. Syst., 11(2)1993

[7]T. von Eicken et al., Active Messages: A Mechanism for Integrated Communication and Computation, in Proc. 19th Ann. Int. Symp. Comput. Architecture, May 1992

[8]Tucker and A. Mainwaring, CMMD: Active Messages on the CM-5, Parallel Computing, 20(4)1994

[9]Oliver A. Mcbryan, An Overview of Message Passing Environments, same to [8]

[10]王鼎兴、庄伟强，一种实现并行计算的新主流技术—NOW，小型微型计算机系统，16(2)1995

[11]Ronald J. Vetter, ATM Concepts, Architectures, and Protocols, CACM, 38(2)1995

[12]MPI Forum, MPI: A Message-Passing Interface Standard, May, 1994

[13]Al Geist, et al., PVM: A Users' Guide and Tutorial for Networked Parallel Computing, 1994