

粗糙集理论及其应用进展

胡可云, 陆玉昌, 石纯一

(清华大学 计算机科学与技术系, 北京 100084)

摘要: 粗糙集理论是一种新型的处理模糊和不确定知识的数学工具。目前已在人工智能、知识与数据发现、模式识别与分类、故障检测等方面得到了广泛应用。首先描述了粗糙集的基本算法及其复杂度, 包括等价关系, 上下近似及各种约简算法; 接着对粗糙集扩展理论, 如可变精度模型, 相似模型等进行了讨论, 然后对粗糙集在数据挖掘、大数据集、粗糙逻辑、多方法融合等领域中的应用进展情况进行了论述, 最后给出了建议的研究方向。

关键词: 粗糙集; 知识发现; 数据分析

中图分类号: TP 18

文献标识码: A

文章编号: 1000-0054(2001) 01-0064-05

Advances in rough set theory and its appliations

HU Keyun, LU Yuchang, SHI Chunyi

(Department of Computer Science and Technology,
Tsinghua University, Beijing 100084, China)

Abstract Rough set theory, a new mathematical tool dealing with vagueness and uncertainty, was introduced by Pawlak in 1982. It has been widely used in the area of AI, data mining, pattern recognition, fault diagnostics, etc. This paper describes the basic algorithms for rough set theory, including equivalent relation, upper/lower approximation and reduction. Then several extensions of rough set theory are discussed such as VPRS, similarity based model, and applications of rough set theory in areas like data mining, rough logic, etc. Further research directions are then discussed.

Key words rough set; knowledge discovery; data analysis

粗糙集 (rough set) 理论是一种新型的处理模糊和不确定知识的数学工具。目前已经在人工智能、知识与数据发现、模式识别与分类、故障检测等方面得到了较为成功的应用。

粗糙集理论具有一些独特的观点。这些观点使得粗糙集特别适合于进行数据分析。如:

知识的粒度性。粗糙集理论认为知识的粒度性

是造成使用已有知识不能精确地表示某些概念的原因。通过引入不可区分关系作为粗糙集理论的基础, 并在此基础上定义了上下近似等概念, 粗糙集理论能够有效地逼近这些概念。

新型成员关系。和模糊集合需要指定成员隶属度不同, 粗糙集的成员是客观计算的, 只和已知数据有关, 从而避免了主观因素的影响。

采用粗糙集理论作为研究知识发现的工具具有许多优点。粗糙集理论将知识定义为不可区分关系的一个族集, 这使得知识具有了一种清晰的数学意义, 并可使用数学方法进行处理。粗糙集理论能够分析隐藏在数据中的事实而不需要关于数据的任何附加信息。

但是, 在粗糙集合应用于实际系统时, 仍然存在一些实际问题。例如约简的有效计算问题, 如何处理数据中的噪音和丢失值问题。为解决上述问题, 有许多工作集中在寻求有效的约简算法和对经典粗糙集理论的扩展上。

1 基本算法及其复杂度

1.1 求等价关系

求等价关系的的最坏复杂度为 $O(|A| |U|^2)$, 其中 A 为属性集合, $|U|$ 为对象集合。这是因为在最坏情况下需要扫描对象集合两次。每个对象一次, 每个对象的等价类一次。一个改进的算法是首先按给定属性集对对象排序, 然后扫描一遍即可。这样它的复杂度就降低到 $O(|A| |U| \lg |U|)$ 。

1.2 上下近似

如果已经给定条件属性和决策属性的等价类

收稿日期: 1999-12-21

基金项目: 国家自然科学基金资助项目 (79990580); 国家“九七三”
基础研究项目 (G1998030414)

作者简介: 胡可云 (1970-), 男 (汉), 湖北, 博士研究生

(或划分),求某个集合的上下近似是直接了当的事情。只需测试该集合在条件属性下的等价类是否包含或部分包含在决策属性下的等价类中即可。等价类的个数在最坏的情况下等于 $|U|$ 。因此在已给定划分(等价类)下求上下近似的复杂度为 $O(|U|)$ 。这样整个的复杂度最坏在 $O(|A||U|\lg|U|)$ 。Chan给出一种在属性增减时上下近似的更新方法^[1]。作者指出该方法的一种可能的应用是动态属性泛化。该方法使得更新后的上下近似可以用原属性集合的上下近似,原属性集合的边界,新增单个属性的上下近似,新增单个属性的边界来表示。这使得增加一个属性时,上下近似的更新只需计算单个属性的上下近似即可。

1.3 属性重要性及核

属性的重要性在于去掉该属性后正区域的变化大小。其时间复杂度的主要取决于正区域。因此计算属性重要性的时间复杂度和计算上下近似的复杂度相同。核的计算可以通过属性的重要性来计算。若属性的重要性大于0,它必是核的成员。所以核的计算可以通过测试 X 中每个属性的重要性来决定。因此计算核的复杂度为 $O(|X||A||U|\lg|U|)$ 。其中核 X 是条件属性的子集

1.4 约简

约简是粗糙集用于数据分析的重要概念。然而最小约简的计算是NP-hard的。因此运用启发信息来简化计算是必要的。事实上,计算最小约简的问题有些类似于机器学习中的最小属性子集问题。如前所述粗糙集中的约简计算可以转化成布尔函数化简问题。因此可以使用许多布尔函数化简中的技巧及算法。有许多作者讨论了计算约简的问题。这里我们介绍一些典型算法

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。则每个主蕴含式均为约简。基本算法可以求出所有的约简,但是只适合于非常小的数据集。基本算法的时间复杂度为 $O(2^{|A|}|A||U|\lg|U|)$ 。

算法2(属性的重要性) 由Hu提出^[2]。该算法非常简单和直观。它使用核作为计算约简的出发点。计算一个最好的或者用户指定的最小约简。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属性,直至该集合是一个约简为止。接着检查该集合中的每个属性,看移走

该属性是否会改变该集合的对决策属性依赖度。如果不影响,则将其删除。此算法的最坏复杂度在 $O(|A|^2|U|\lg|U|)$ 。因为循环的执行次数最多为 $|A|$,而求属性间的依赖程度的复杂度和计算正区域相同。

算法3(遗传算法) 已经有不少用遗传算法计算约简的算法。各种算法的不同之处主要在表示和适值函数的不同。这里介绍具有代表性的Bjorvand和Komorowski提出的遗传算法^[3]。表示:每个位串代表区分矩阵的一项,即两个对象的区分属性集。某位为1时表示该属性存在,否则不存在。这样每个位串是一个约简的候选。定义适值函数如下

$$F(v) = \frac{N - L_v}{n} + \frac{C_v}{(m^2 - m)/2}$$

其中: N 是属性集合的长度, L_v 是 v 中1的个数。 C_v 是 v 能区分的对象组合的个数。 m 是对象的个数。该函数由两部分组成。前一部分的目的是希望 L_v 的长度尽可能得小。后一部分希望区分的对象尽可能多。在设计初始种群时,可以考虑将核或专家认为必要的属性加入种群中,以加快算法的收敛。

算法4(复合系统的约简) Kryszkiewicz和Rybinski研究了在复合信息系统中寻求约简的问题。即怎样利用现有的子系统的约简求复合系统的约简^[4]。其主要思想是将布尔函数的化简问题转化成集合空间中的边界搜索问题。而在已知子系统的约简的情况下,复合系统的搜索空间将得到简化。设有信息系统 S_1, S_2 。它们的属性集合相同。设 f_1 和 f_2 分别是它们的区分函数。则整个信息系统 S 的区分函数 f 可表示为 $f = f_1 \wedge f_2 \wedge f_{12}$ 。其中 f_{12} 代表 S_1, S_2 中的对象分别作为纵横坐标组成的区分函数。根据上面的讨论,如果我们已知 S_1 和 S_2 的约简时,则 S 的约简只需在空间 $[MINS(f_1 \wedge f_2), \{A\}]$ 上搜索而不必从头开始。其中 $MINS(f_1 \wedge f_2)$ 是两个子系统的约简的并的最小值。因而搜索空间大大减小。

算法5(扩展法则) Starzyk, Nelson和Sturtz提出一种新概念,称为强等价(strong equivalence),进而发展为扩展法则,用于快速简化区分函数^[5]。两个属性称为局部强等价,若它们在区分函数的所有项中同时出现或不出现。当两个属性是局部强等价时,它们就可以仅用一个属性代替。实验表明该算法比基本算法快数十到数百倍。因而能处理更大的数据集。

算法6(动态约简) 动态约简在某种意义上是

给定决策表中最稳定的约简,它们是在从给定决策表中随机抽样形成的子表中最常出现的约简。动态约简能够有效的增强约简的抗噪音能力^[6]。动态约简的计算过程较为简明,主要是对决策表进行采样,然后对采样后的决策表计算所有约简。在所有的子表中保持不变或近似保持不变的约简就是动态约简。

2 扩展模型

粗糙集理论应用于数据分析时,会遇到噪音、数据缺失、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。

2.1 可变精度模型(VPRS)

在数据集中存在噪音等干扰情况下,经典理论会由于对数据的过拟合而使其对新对象的预测能力大为降低。而在实际应用中,噪音是在所难免的。为增强粗糙集合模型的抗干扰能力,Ziarko提出了一种可变精度RS模型^[7]。该模型通过引入一个精度,从而具有一定的容错性。

VPRS的主要扩充体现在它允许一定的误分类率。为此,引进一个精度 $U(0 \leq U \leq 0.5)$ 。定义 U 多数包含关系为:若把集合 X 中的元素分类到集合 Y 中,则会犯分类错误的可能性小于 U 。显然,若 $U=0$,则 U 多数包含关系就退化成标准的包含关系。这样就可以用 U 多数包含关系来解释粗糙集中的上下近似。

集合 X 的下近似可解释为那些 U 中的元素分类到 X 中分类错误率不大于 U 的等价类集合。 X 的 U -上近似 $R_U X$ 包含所有 U 中不可能在误分类率小于 U 之下分类到 $-X$ 中的所有元素。

根据经典粗糙集同样的方式,利用集合的下近似,可以定义属性的 U -近似依赖。粗糙依赖量度是精确分类,即对对象的无错误分类的综合能力的评价,而近似依赖量度是度量分类错误率在预先给定的可容忍极限 U 之内的分类能力。近似分类与粗糙分类相反,不能解释为属性的函数或部分函数依赖。近似依赖的性质弱于函数依赖的性质,例如传递性就不再成立。

根据扩展的近似依赖的定义, U 约简或称近似约简可以定义为在近似依赖度不变的前提下的最小子集。VPRS模型是和经典粗糙集兼容的。因为只要令 $U=0$,那么就与经典模型一致了。因此,VPRS模

型能够保持绝大多数经典模型的良好性质。这为它的广泛应用打下了基础。

Katzberg和 Ziarko进一步提出了不对称边界的VPRS模型,即在上下近似的定义中的 U 可以是不相同的。从而使此模型更加一般化^[8]。

2.2 相似模型

经典粗糙集模型的基础——不可区分关系——是很强的。在数据中存在缺失的属性值的时候(在数据库中很普遍),不可区分关系或者说是等价关系无法应付这种情形。为扩展粗糙集的能力,有许多作者提出了用相似关系来代替不可区分关系作为粗糙集的基础^[9]。

在使用相似关系代替粗糙集合中的不可区分关系后,最主要的变化就是相似类不再形成对原集合的划分了。它们之间是相互重叠的。类似于等价类,可以定义相似集。即所有和某个元素 x 在属性集合 B 上相似的集合 $SIM_B(x)$ 。值得注意的是 $SIM_B(x)$ 中的元素不一定属于同一决策类。因此还需定义相似决策类,即相似集对应的决策类集合。

由于相似集的元素并不一定属于同一个决策类,为此定义相对吸收集。子集 $Y \subseteq U$ 称为相对吸收集,如果对于每个 $x \in U$,存在 $y \in Y$ 与之相似,并且具有同样的决策值。显然相对吸收集可以用来进行数据削减。利用相似集可以很容易地定义正区域的概念。它就是所有包含在决策类中的相似集的并。依赖度和约简的概念都可以类似经典集合的方式定义。

实践证明,相似模型在实践使用中具有比经典粗糙集模型更好的性能。在解决数据库中缺少值的情况时,一个简单的相似关系可以定义为(其中*代表不知道或不关心):

$$\begin{aligned} \mathfrak{f}(x, y) &= \{x \in U, y \in U \mid \forall a \in C, \\ a(x) &= a(y) \text{ or } a(x) = * \text{ or } a(y) = *\}. \end{aligned}$$

3 应用

可以说,粗糙集理论由于其数据挖掘方面的应用而受到广泛的关注。最近几年,粗糙集理论的应用研究得到了长足发展。这里从几个方面简述相关的应用。由于在粗糙集的文献非常多,我们仅列举其中极少数代表性应用。

数据缩减与规则生成。Kohavi和 Frasca等用实验证明,数据库中最有用的子集并不一定是粗糙集中的相对核,甚至可能不包括完全的核属性集^[10]。Ning Shan则讨论了基于RS的从数据中发

现规则的增量自适应算法^[11]。提出了一种找到最大泛化的规则和约简的算法和规则与缩减的增量算法。Grzymala-Busse和 Zoubi比较了同时使用可能规则及确定规则和只使用确定规则的性能,发现前者产生较小的错误率^[12]。Choubey和 Deogun等的研究得出了同样的结论。他们还在属性选择的题目下研究了近似约简问题,并给出了几个启发式算法^[13]。Lenarcik和 Piasta研究了在每个对象的Cost不一样时的粗糙分类器。他们的主要方法是对所有的对象定义一个新的Cost属性^[14]。Mollesstad T和 Komorowski J提出了在粗糙集框架下缺省规则生成的格搜索算法^[15],并给出一组启发式搜索策略。

大数据集。由于粗糙集在数据挖掘中具有较大的计算复杂度,受关联规则挖掘算法的启发,有些作者提出了将关联规则挖掘技巧应用于粗糙集的确定和可能规则生成中来,以减小粗糙集方法的计算复杂度^[16]。Nguyen和 Skowron等描述了一种决策表分解方法^[17]。他们首先使用遗传算法在决策表中寻找代表性的模板(类似如一条支持度最大的规则),然后将决策表一分为二。满足模板的为一个部分,不满足的为另一部分。将该过程递归进行,直至决策表的大小满足要求为止。然后再对小决策表生成规则。当新对象到来时,从顶部开始匹配,直至叶子的规则。

多方法融合。Jelonek等研究了将RS理论用于神经网络训练数据的预处理,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,并且保持了较低的稳定的近似分类误差率^[18]。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^[2],首先使用面向属性的概念树爬升技术对属性进行泛化,然后使用RS方法计算缩减并生成规则。由于在泛化过程中消除了不必要的属性值和在缩减过程中去掉了不相关的属性,最后的规则是很一般的形式并且可用高层次抽象概念表达。Lingras和 Davies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^[19]。在该算法中,基因用粗糙数表示。

信息检索。Beaubouef等在RS理论上提出了一种Rough关系数据库模型^[20],并定义了各种Rough关系算子。该模型将RS的重要性质引入到基本关系模型中,从而使之具有更好的检索能力和适应性。在此模型中,查询结果返回的是基于属性的Rough关系,它不仅包含一个查询的确定应答,还

包含可能的应答,例如上近似所包含的元组等。

粗糙逻辑。不少逻辑学家和理论计算机科学家试图通过RS建立Rough逻辑。Pawlak等在CACM 1995年11期上发表综述,认为研究粗糙逻辑——基于RS的不精确推理逻辑——是粗糙集应用研究中最重要课题之一。Lin和Liu等基于拓扑学观念定义了Rough下近似算子 L 和Rough上近似算子 H ^[21],这2个算子的语法性质分别与模态逻辑中的必然算子(和可能算子 \Diamond 十分相似,因而带 L 和 H 算子的逻辑公式被称为Rough逻辑公式,并且建立了与模态逻辑相似的公理化Rough集的逻辑演绎系统和相平行的演绎规则。Yao和Lin通过研究粗糙集和模态逻辑的关系,提出并研究了一系列扩展粗糙集的性质^[22]。通过使用不同的二元关系作为粗糙集的基础,可以导出不同的粗糙集代数模型,相应为不同的模态逻辑。

决策支持。决策分析不仅仅是分类任务。它还要求对属性的评价的标准。例如同类的商品我们认为低价比高价好等等。为使经典粗糙集理论适用决策支持的要求, Greco, Matarazzo和 Slowinski给出了扩展经典粗糙集的方法^[23]。主要是采用一种和评价准则有关的支配关系(dominance,自反和传递的)代替等价关系的方法。并提出以模糊属性评估方法代替粗糙集中的属性重要性方法。

原型系统。典型系统如KDD-R, LERS, Rough DA& RoughCLASS, Rosetta, Rough Enough, Grobian等^[24]。这些系统为了实际的需要,一般会对经典理论有所扩展。

其它方面的应用例子有字符识别,医疗诊断,市场预测等等。这些应用主要是利用粗糙集及其扩展方法进行规则获取后在具体领域的应用。这里我们不一一讨论。有兴趣的读者可参看有关文献。

4 结 语

粗糙集合以其独特的优势正在赢得越来越多的研究者关注,并在各个应用领域获得了广泛的应用。然而这仍是一个极其年轻并在高速发展的学科。下面从数据库知识发现角度列举一些可能的研究方向及应用领域。这些方向实际上是相互关联的。

高效约简算法。高效的约简算法是粗糙集应用于知识发现的基础。现尚不存在一种非常有效的方法。因此寻求快速的约简算法及其增量版本仍然是主要研究方向之一。

大数据集问题。现实中的数据库已经越来越大。

粗糙集理论如何应付这一挑战仍旧是一个问题。虽然现在已经有一些有益的探索,但是还没有找到一种令人满意的方法。可能的解决方案有采样、并行化等。需要发展相应的算法。

多方法融合。现在有许多种数据挖掘方法。实验表明,还没有一种在所有的测试集上表现都比其它方法出色。因此多种方法的融合可能是进一步提高分类效能的方法之一。此外,最近提出的 committee 学习方法在粗糙集方法上的应用也是有意义的课题之一。

参考文献 (References)

- [1] Chan C C. A rough set approach to attribute generalization in data mining [J]. *Information Sciences*, 1998, 107: 169–176.
- [2] Hu X. Knowledge Discovery in Databases An Attribute-oriented Rough Set Approach [D]. University of Regina, Canada, 1995.
- [3] Bjorvand A T. ‘Rough Enough’—A system supporting the Rough Sets Approach[EB/OL]. <http://home.sn.no/~toivill>.
- [4] Kryszkiewicz M, Rybinski H. Finding reducts in composed information systems [A]. Ziarko W P (eds). *Proceedings of RSKD 93* [C]. London: Springer-Verlag, 1994. 261–273.
- [5] Starzyk J, Nelson D E, Sturtz K. Reduct generation in information system [J]. *Bulletin of International Rough Set Society*, 1999, 3(1/2): 19–22.
- [6] Bazan J G, Skowron A, Synak P. Dynamic reducts as a tool for extracting laws from decisions tables [A]. Ras Z W, Zemankiva M. *Methodologies for Intelligent Systems* [C]. Berlin: Springer-Verlag, 1994. 346–355.
- [7] Ziarko W. Variable precision rough set model [J]. *Journal of Computer and System Sciences*, 1993, 46: 39–59.
- [8] Katzberg J D, Ziarko W. Variable precision rough sets with asymmetric bounds [A]. Ziarko W P. *Proceedings of RSKD 93* [C]. Springer-Verlag, 1994. 167–177.
- [9] Kryszkiewicz M. Rough set approach to incomplete information systems [J]. *Information Sciences*, 1998, 112: 39–49.
- [10] Kohavi R, Frasca B. Useful feature subsets and rough set reducts [A]. *Proc on RSSC 94* [C]. 1994.
- [11] Shan N, Ziarko W. Data-based acquisition and incremental modification of classification rules [J]. *Computational Intelligence*, 1995, 11: 357–370.
- [12] Grzymala-Busse J W, ZOU X. Classification strategies using certain and possible rules [A]. *Proc of 1st Conf on RSCTC* [C]. Poland, 1998. 37–44.
- [13] Choubey S K, Deogun J S, Raghavan V V, et al. A comparison of feature selection algorithm in the context of rough classifiers [A]. *5th IEEE Int Conf Fuzzy Systems* [C]. New Orleans, 1996. 1122–1128.
- [14] Lenarcik A, Piasta Z. Rough classifiers sensitive to costs varying from object to object [A]. *Proc of 1st Conf on RSCTC* [C]. Poland, 1998. 222–230.
- [15] Mollestad T, Komorowski J. A rough set framework for mining propositional default rules [A]. Pal S K, Skowron A eds. *Rough Fuzzy Hybridization* [C]. Springer, 1999. 233–262.
- [16] Lin T Y. Rough set theory in very large databases [A]. *Proceedings of IMACS Multiconference* [C]. volume 2, Lille, International Association for Mathematics and Computers in Simulation, 1996. 942–947.
- [17] Nguyen S H, Skowron A, Synak P, et al. Knowledge discovery in data bases: Rough set approach [A]. *Proc of IFSA 97* [C]. Prague, 1997. 204–209.
- [18] Jelonek J, Krawiec K, Slowinski R. Rough set reduction of attributes and their domains for neural networks [J]. *Computational Intelligence*, 1995, 11(2): 339–347.
- [19] Lingras P, Davies C. Rough Genetic algorithms [A]. Zhong N, Skowron A eds. *Proc 7th Intl Wksp on RSFD* [C]. Springer, 1999. 38–46.
- [20] Beaubouef T, Petry F E. A rough set model for relational databases [A]. Ziarko W P eds. *Proc of RSFD 93* [C]. Springer-Verlag, 1994. 100–107.
- [21] Lin T Y, Liu Q, Yao Y Y. Logic systems for approximate reasoning via rough sets and topology [A]. Ras Z W, Zemankiva M eds. *Methodologies for Intelligent Systems* [C]. Berlin: Springer-Verlag, 1994. 65–74.
- [22] Yao Y Y, Lin T Y. Generalization of rough sets using modal logics [J]. *Intelligent Automation and Soft Computing*, 1996, 2(2): 103–120.
- [23] Greco S, Matarazzo B, Slowinski R. Fuzzy measure technique for rough set analysis [A]. *6th Euro Congress Intelligent Techniques & soft Computing* [C]. Germany, 1998. 99–103.
- [24] Rough set based kdd Systems[EB/OL]. <http://www.kdnuggets.com>, 1999.