粗糙集理论分析及其应用研究

覃 宝 灵
（佛山科学技术学院信息与教育技术中心，广东 佛山 528000）

摘 要：本文阐述粗糙集理论的基本概念，探讨粗糙集理论中知识约简和规则提取的重要性，通过分析、比较，把这些理论和技术应用于实际中，取得了显著的效果，对其在信息系统中的应用具有一定的研究价值。

关键词：粗糙集；知识约简；规则提取；遗传算法

1、前言

随着信息技术的飞速发展和广泛应用，面对信息系统中不完整、不精确或不确定的数据如何有效分析处理？如何发现隐藏在信息系统中的有用知识和潜在的规律？为了解决这些问题，学术界和研究者们采用了粗糙集理论。粗糙集理论是由波兰数学家 Z.Pawlak 在 1982 年提出的[1]，它是一种分析处理不完整性、不精确性、不确定性知识的数学工具。该理论不需要任何初始或附加信息，直接利用已知的知识库，将知识库中的不确定或不精确的知识进行近似的划分，并对所划分的知识域确定其支持程度。

目前，该理论已成为信息科学和认识科学领域的研究热点之一，随着研究的深入，该理论得到了很大的发展和壮大，并已成功应用于人工智能、模式识别与分类、知识发现与决策分析、专家系统、数据挖掘、故障检测、金融、医学、生物学等领域。

2、粗糙集的基本理论定义

粗糙集理论是一种研究不完整、不确定性知识的数学工具[2]。在信息系统中，对知识的理解和表示是人们首先思考的问题，同时也是比较难解决的问题，从目前研究来看，对这些问题的解决，粗糙集理论和技术是比较理想的方法。

定义 1：（信息系统）设一个信息系统[3]S＝（U，A，V，f）,这里，

①U 是对象的非空有限集合，即称为论域，记为：U＝{x_1,x_2,---,x_n}；

②A 是属性的非空有限集合，记为：A＝{A_1，A_2，---，A_m}；

③V 是属性的值域集，记为：V＝{V_1，V_2，---，V_m}，且 V_i 是属性 A_i 的值域；

④f 是信息函数，即 f：U×A→V，$f(x_i, A_j) \in V_j$。

在信息系统中，若属性集合 A 由条件属性集合 C 和决策属性集合 D 组成，且 C ∪ D ＝ A，C ∩ D ＝ Φ，则称 S 为决策系统，又称决策表。

定义 2：（等价关系）设知识表示系统S=(U,A ,V,f),若属性集合 $p \subseteq A$ 时,称P的不可分辨关系Ind(P)是U上的等价关系,其中Ind(P)={(x,y)∈U × U｜$\forall a \in p$,f(x,a)=f(y,a)}。由Ind(P)导出的所有等价类集合记为U／P，它构成了论域的一个划分，含有元素x的等价类,记为$[x]_p$ 。

定义3：（下近似、上近似、边界域）设 $X \subseteq U$ 是一个集合，R是一个定义在U上的等价关系。有：

①若$\underline{R}(X)$=U{Y∈U／R：Y∈X}，则称$\underline{R}(X)$为X的R下近似集；

②若$\overline{R}(X)$=U{Y∈U／R：Y∩X≠Φ}，则称$\overline{R}(X)$为X的R上近似集；

③若R(X)=$\overline{R}(X)$—$\underline{R}(X)$，则称R(X)为集合X的边界域。若R(X)是空集，则称集合X关

于集合R是清晰的；反之，称集合X为关于集合R的粗糙集。

定义4：设R是一族等价关系，且{R}∈R，若Ind(R)≠Ind(R-{R})，则称{R}为R中不可省略的，否则称{R}为R中可省略的。当每一个{R}都是R中不可省略的，则称{R}为独立的。

定义5：设P⊆R，当P为独立的，且Ind(P)=Ind(R)，则称P是R的一个约简，记为Red。R中所有不可省略关系构成的集合称为R的核，记为Core。可推并证得：Core=∩Red。

3、粗糙集理论的知识约简及规则提取分析

在数据挖掘过程中，粗糙集理论的核心是知识约简，其算法是在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。其知识处理模型如图 3-1 所示：

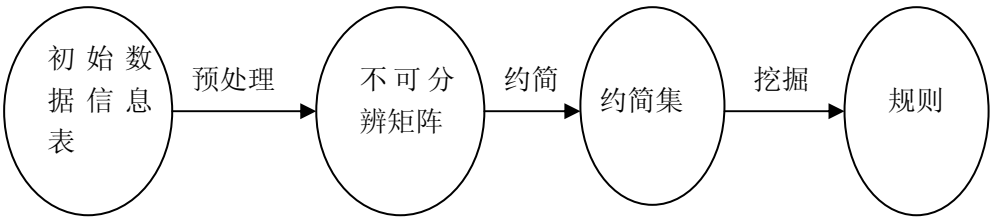

图 3-1

3.1 知识约简分析

在粗糙集理论中，“知识”理解为一种分类能力，即对数据的划分，可用集合表示，例如，假设给定数据集U和等价关系集P，若用P来划分U，则称其为知识。知识约简是指在保持知识库的分类或决策能力不变的条件下，删除其中不相关或不重要的知识，从而可以简化判断规则，提高决策效率。

在实际应用中，通常用决策表来描述论域中的每个对象，其实，它是二维表，每一行表示一个对象，每一列表示对象的每一种属性，而属性又分条件属性和决策属性两类，在论域中的对象，根据条件属性的不同，被划分到具有不同决策属性的决策类中。由于一个属性对应一个等价关系，一个表可以表示一族等价关系，即知识库，所以知识约简可以转化为属性约简。

在决策表分类过程中，可能存在多个约简，若将这些约简交集，则称其为核，它是计算所有约简的基础，是知识最重要部分的集合，对知识约简时不能删除它。核的属性是分类的关键属性，所以在信息系统中，如何计算出核的属性呢？通常的算法是首先删除重复的实例和不关联的属性，其次删除每个实例的多余属性，最后计算出最小约简，并通过最小约简，求出逻辑规则。当前随着知识库的不断扩大，知识约简的复杂性也越来越大，要计算所有约简和最佳约简，可以说是很大的难题。通过分析研究，目前比较理想的方法是用遗传算法来求较优的知识约简。遗传算法流程图如图 3-2 所示：

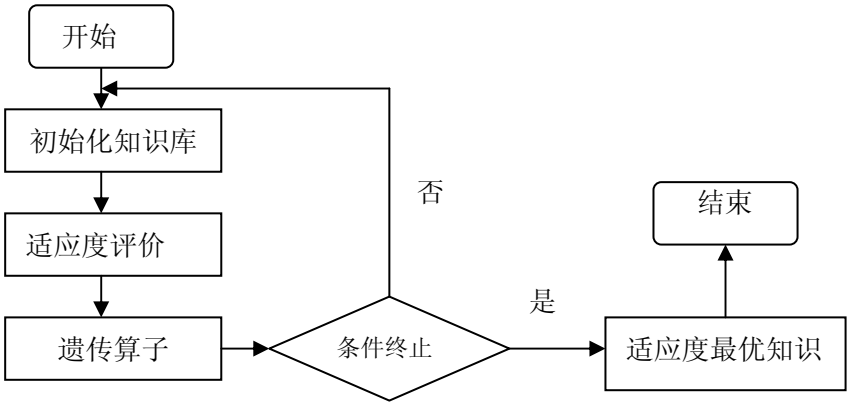

图 3-2

3.2 规则提取分析

知识约简最终目的不仅仅是为了减少知识库中的条件属性，压缩数据量，而是为了知识发现打好基础。知识发现的任务是从知识库中找出知识之间的内在联系，并提取决策规则。

规则提取其实就是决策规则的约简结果，主要对条件属性值的约简。在决策表中，每一行对应一条决策规则，首先计算决策规则的条件属性的核值，其次求它的条件属性值的约简。如果约简表中出现重复行，则将其删除，因为其表示同一决策规则。属性值的约简需要对决策表进行多次遍历，计算量较大，并且存在属性冗余和规则冗余。由于最优的决策树已经被证明是一个 NP 难题[4]，所以，目前比较理想的方法是用更优的启发式函数来构造决策树，并提取决策规则。其方法为以属性重要性评价指标作为信息熵函数，对条件属性进行选择，对属性间的依赖性和冗余性进行充分考虑，对不相容决策表进行正确分类，从而弥补了 ID3 算法对属性间依赖性强调不够的缺点，解决了决策树中子树的重复和有些属性在同一决策树上被多次选择的问题。

3.3 问题存在分析

尽管目前粗糙集理论研究取得令人瞩目的成果，但仍然存在一些至今还没有很好解决的问题。

（1）分类在实际应用中存在局限性问题[5]。由于粗糙集理论处理的分类是精确的，只考虑数据集的完全“包含”与“不包含”，不存在含糊，还有它处理的的对象是已知的，并且从模型中提取的结论仅适用于这些对象。

（2）属性约简有效性计算问题。在知识约简的过程中，如何处理数据中的噪音和丢失值问题，连续属性离散化问题等，目前还没有找到令人满意的约简算法来处理这些问题。

（3）规则提取不稳定性问题。由于粗糙集理论处理的错误判断的决定性机制非常简单，所以，由粗糙集产生的决策规则很不稳定而且有较差的分类精确性[6]。因些为了解决这些问题，粗糙集理论必须与其它数据挖掘方法结合，例如：模糊集、神经网络等。

除以上问题之外，还有例如：最优间隔断点选取及知识获取等技术问题。

4、粗糙集理论的应用

虽然粗糙集理论发展只有二十多年，但随着信息技术的飞速发展和广泛应用的引领下，其在各领域的研究成果是令人瞩目的，它的生命力在于它具有较强的实用性。

4.1 神经网络中的应用

在专家系统中，由于知识获取是非常关键的阶段，同时对它们的定义又是很困难，为了有效解决这些问题，由苏丹卡同大学、马来西亚大学和普恰大学的M. E. Yahia、R. Mahmod等研制的粗糙神经专家系统中提出将神经网络作为专家知识库[7]。主要是运用粗糙集作为数学工具来处理不确定与不精确数据，将粗糙集和神经网络结合形成一种新的知识库结构，它是基于粗糙分析约简和神经网络的结合上，这样结合有利于发挥各自的特点，达到优势互补的效果。目前该系统应用于医学诊断中肝炎病例的检测。

4.2 证券数据分析中的应用

Golan和Ziarko应用粗糙集理论分析了十年间股票的历史数据，研究了股票价格与经济指数之间的依赖关系，获得的预测规则得到了华尔街证券交易专家的认可。

4.3 知识发现中的应用

目前，粗糙集理论在 KDD（Knowledge Discovery in Database）的应用得到了飞速的发展，基于粗糙集理论的知识发现方法是 KDD 的重要方法之一。Hu,X.等人将基于属性的归纳学习方法和基于粗糙集的机器学习方法相结合，建立了数据库学习系统。Beaubouef,T.等人将粗糙集算法嵌入数据管理系统中，提出了粗糙关系数据库模型，使查询处理更灵活，大大增强了系统的检索能力。

4.4 决策分析中的应用

基于粗糙集理论的决策规则是在分析经验数据的基础上得到的，它允许决策对象存在一些不太明确的属性。在决策分析中，意大利学者Salvatore Greco和波兰学者Roman Slowinshi提出将粗糙集应用于多标准决策分析[8]。又例如希腊发展银行ETEVA应用粗糙集理论协助制订信贷政策，是粗糙集理论多准则决策方法的一个成功范例。

4.5 故障诊断中的应用

在故障诊断方面，由于故障发生的复杂性、不确定性及对故障原因认识的局限性，所以对相关故障信息的分析和处理往往存在一定的困难，用粗糙集方法约简去冗余信息，同时结合其它软计算方法对生成的决策规则进行故障判断，从而提高系统的诊断效率[9]。

4.6 医疗诊断中的应用

在医疗诊断方面，用粗糙集方法根据以往病例归纳出诊断规则，用来指导新的病例。人工预测早产准确率只有17%-38%，应用粗糙集理论可提高到68%～90%。

除了以上例举的应用领域之外，还有：软件工程数据分析、近似推理、粗糙控制、图象处理、地震预报、电力系统分析等，其应用空间很广阔，是目前信息科学的研究热点。

5、结束与展望

粗糙集理论虽然是一门十分年轻的学科，但它目前已成为一个前沿研究领域，由于它是一种有效处理不确定性知识的数学工具，同时在智能领域得到广泛应用，所以被公认为人工智能领域最具潜力的技术之一。从国内外研究资料看，粗糙集理论还处在继续发展之中，理论上的一些问题还需要解决，但我们相信，随着科学的进步和研究及探讨的深入，存在的问题会得到解决的，同时更坚信粗糙集理论在信息系统中更具有广阔的发展空间和实用价值。

参考文献：

[1]Pawlak Z.Rough[J]. International Journal of Computer and Information Sciences, 1982,11(2):341-358.

[2]Pawlak Z.Rough sets: theoretical aspects of reasoning about data[M]. Boston:Kluwer Academic Publishers, 1991:65-90.

[3]毛国君，段立娟，王实，石云等编著．数据挖掘原理与算法[M]．北京：清华大学出版社，2005：26-28.

[4]洪家荣,丁明峰,李星原等.一种新的决策树归纳学习算法[J].计算机学报,1995,18(6):470-474.

[5]蒋良孝，蔡之华，刘钊.一种基于粗糙集的决策规则挖掘算法[J].微型机与应用，2004（3）：7-8.

[6]Supriya K D，Krishna P R．Clustering web Transactions Using Rough Approximation[J]．Fuzzy Sets and Systems，2004(148)：130-139.

[7] Yahia M，Mahmodr R，Sulmmann N．Rough neural expert systerm[J]．Expert system with Applicati0Ils，2002，18：87-99

[8]Alvatore G，Bentto M，Roman S．Rough set theory for multi criteria decision analysys[J]. European Journal of Operational Research, 2001,129:1-46

[9]李千目，戚勇，张宏，等.基于粗糙集神经网络的网络故障诊断新方法[J].计算机研究与发展，2004，41（10）：1696-1702