不完备信息系统知识获取的粗糙集理论与方法

周献中 黄 兵 李华雄 魏大宽 著

南京大学出版社

不完备信息系统知识获取的粗糙集理论与方法

○周献中 黄 兵 李华雄 魏大宽 著

南京大学出版社

图书在版编目(CIP)数据

不完备信息系统知识获取的粗糙集理论与方法 / 周献中等著. — 南京 : 南京大学出版社, 2010.12
ISBN 978-7-305-07887-3

Ⅰ. ①不… Ⅱ. ①周… Ⅲ. ①信息系统—数值计算—研究 Ⅳ. ①TP18

中国版本图书馆 CIP 数据核字(2010)第 235266 号

内容简介

如何从海量的不完备、不确定数据中发现有用的知识结构是智能科学研究的热点与难点. 粗糙集理论因其在数据处理技术和方法方面的特点和优势而被国内外相关领域的学者广泛关注,且在针对不完备数据的知识获取方面取得了一定的进展. 本书就是以此为背景,以不完备信息系统为对象,以粗糙集为工具,系统介绍基于粗糙集的不完备信息系统知识获取相关理论和方法研究成果,侧重于不完备信息系统的粗糙集模型拓展、知识约简理论及规则提取算法等. 本书主要内容是作者承担国家自然科学基金相关项目研究成果的系统性反映.

本书可供信息与计算机科学、控制科学与工程、管理科学与工程、应用数学等专业的大学高年级学生、研究生、高校教师以及相关科技人员阅读和参考.

出版发行 南京大学出版社
社　　址 南京市汉口路 22 号　　邮　　编 210093
网　　址 http://www.NjupCo.com
出 版 人 左　健

书　　名 不完备信息系统知识获取的粗糙集理论与方法
著　　者 周献中　黄　兵　李华雄　魏大宽
责任编辑 胥橙庭　　编辑热线 025-83686308

照　　排 南京南琳图文制作有限公司
印　　刷 南京紫藤制版印务中心
开　　本 787×960　1/16　印张 13　字数 238 千
版　　次 2010 年 12 月第 1 版　　2010 年 12 月第 1 次印刷
ISBN 978-7-305-07887-3
定　　价 38.00 元

发行热线 025-83594756
电子邮箱 Press@NjupCo.com
Sales@NjupCo.com(市场部)

前　言

21世纪是知识经济与信息技术引领潮流的时代. 当今信息科学新技术正以几何级数的方式不断更新与高速发展,人类社会已处在一个信息急剧膨胀的时代. 现代通信技术、计算机技术和网络技术正日益充斥我们生活的每一个角落,并将从根本上改变人们的生活方式. 信息技术,尤其是网络技术的快速发展使得人们可以随时随地快速获取大量的数据,由此积累的数据越来越多. 在这些海量的数据背后隐藏着许多重要的信息,人们希望能够对其进行多方面、多层次的分析,以便能更好地利用这些数据,为进一步的决策提供依据.

普通的数据库系统虽然可以高效地实现数据的录入、查询、统计等功能,但无法发现数据中存在的关系和规则,无法根据现有的数据预测未来的发展趋势. 人们虽然可以轻而易举地获取大量的数据,但在如何处理这些数据以发现数据背后隐藏的知识方面却陷入了困境,迫切需要发展系统的方法解决该问题. 在这样的背景之下,20世纪末人工智能研究中的一个新领域——数据挖掘(Data Mining,DM)和数据库知识发现(Knowledge Discovery in Database,KDD)逐渐引起人们的关注并开始快速发展. DM与KDD通过设计特定的知识结构和数据处理算法,借助具有快速计算能力的计算机,有效帮助人们从海量数据中提取有用信息,将数据资源转换为有用的信息和知识资源,进而帮助人们科学地做出各种决策.

另一方面,人们除了要面对海量数据外,同时还可能面临数据质量低下的问题,即丰富的不确定信息. 事实上,现实世界中客观事物或现象往往是不确定的,或称为具有不确定性、不完备性,同样,人们主观的各个认识领域中的信息和知识大多也是不精确的. 这种知识、信息的不确定性就要求在对它们进行表示和处理时能够反映这种不确定性. 因此,如何表示和处理知识的不确定性就成为重要的研究领域.

关于信息不确定有如下几种涵义：(1) 信息不准确；(2) 信息不完备；(3) 模糊信息；(4) 不确定性假设. 处理不确定信息有很多方法，如证据理论、模糊理论、概率统计等. 但这些理论需要数据集合的额外信息，如证据理论中的信任函数、模糊理论中的隶属函数、概率统计中的分布函数等. 可想而知，通过不确定海量数据库去确定合适的信任函数、隶属函数或概率分布是不现实的.

粗糙集理论(Rough Set Theory，RST)正是这样一种既能满足数据集合不同简洁程度表示的要求，又不需要数据额外信息的处理不确定信息的知识表达、归纳和推理的数学理论. 它能在保持原数据集合分类能力或决策能力不变的前提下消除冗余的信息，从而获得知识的简洁表达. 它最突出的优点是“让数据自己说话”，即不需要数据集合之外的任何信息. 因此，利用 RST 获得的知识更具客观性.

基于粗糙集理论的 KDD 研究对象是信息系统(由对象集、属性集、属性值构成的二维数据表)，最基本的概念是不可分辨关系(等价关系)，即由相同信息所标识的对象是不可分辨的. 在此基础上，引入上近似集和下近似集分别刻画知识的可能性和确定性，引入边界集来描述知识的不确定性与模糊性，引入约简与求核等方法对知识进行约简. 因而，知识约简便成为利用 RST 进行数据分析的基本任务. 传统 RST 研究的对象是没有数据缺省的完备信息系统，其知识约简是在基于不可分辨关系的前提下确保分类不变而消除冗余属性.

在现实世界中，由于各种各样的原因，常常使得系统中某些数据缺省(失)，即存在不完备的信息系统(Incomplete Information System，IIS). 此时必须要对 RST 进行扩充. 目前，针对 IIS，RST 扩充主要有两类方法：一类是通过领域专家把所缺的数据补齐，间接地将不完备信息系统转化为传统的 RST 能够处理的完备信息系统(称为间接法)；另一类是直接把 RST 中的相关概念在 IIS 中进行适当扩充(称为直接法). 直接法较之于间接法而言，没有领域专家主观因素的参与，更具有客观性，因此引起了许多学者的兴趣. 例如，Kryszkiewicz 提出了基于相容关系(也称为容差关系)的粗糙集扩展模型；

Stefanowski等人提出了基于非对称相似关系和量化容差关系的粗糙集扩展模型；Yao通过定义邻域算子提出了一般二元关系下的粗糙集模型；Greco等人提出了基于优势关系的粗糙集模型；王国胤提出了基于限制容差关系的粗糙集扩展模型等等. 目前，将经典粗糙集模型中的等价关系拓展为较为宽松的二元关系是研究粗糙集模型在IIS中应用的主流方法之一.

本书即以此为主线，在总结已有IIS粗糙集模型及其知识约简研究的基础上，主要介绍具有原创性的联系度、基于γ-相容关系等概念的多种IIS的粗糙集拓展模型，并系统研究其相应的约简方法. 同时，我们通过引入相容矩阵与相似矩阵，对基于相容关系与相似关系的粗糙集拓展模型作了进一步探讨，研究了相应的粗计算、属性约简以及决策规则提取的矩阵算法问题. 此外，在IIS中，还存在这样一类问题：不仅系统中的数据有缺省的，而且决策目标是模糊的情形. 这类问题在现实中更加普遍. 为此，人们既需要对海量数据进行有效地理解和分析，同时要对不完备信息进行操作处理，还要考虑对模糊决策的要求. 我们把它称为不完备模糊决策信息系统，本书针对这一类信息系统的粗糙集模型和约简理论也展开了深入的研究. 最后，作为补充，还给出了一种基于区间集的IIS规则提取方法，利用新的数学工具对IIS的知识获取方法作了进一步探讨.

本书力图通过对上述研究成果的系统整理和详细介绍，为读者解决同类问题提供一种可参考或借鉴的新视角与方法，并希望能对丰富和发展粗糙集理论、促进粗糙集方法在智能科学与知识工程中的应用做出微薄的贡献. 因此，本书的读者群主要定位在具有代数学基础的高年级大学生、研究生、高校教师和相关的科技人员.

借本书出版的机会，作者衷心地感谢国家自然科学基金委员会多年来对本研究给予的支持(国家自然科学基金资助项目号：70571032，90718036，70971062)；感谢南京大学出版社及薛志红副总编对本书出版的支持和帮助；感谢为本书有关内容及成果做过贡献的研究生们，他们是郭玲博士、何新博士、赵亚琴博士、张蓉蓉硕士、陆琦硕士、朱梅

梅硕士、李友江硕士等；感谢多年来关心作者并为本书提出中肯建议的同行们！

由于作者的水平有限，书中定有不少错误与不妥之处，热忱欢迎广大读者批评指正！

作　者

2010 年 5 月　南京大学

目　录

第 1 章　基本概念

1.1　引　言

粗糙集理论作为一种处理不精确、不确定和不协调数据的非经典的数学理论，是由波兰科学院院士、数学家 Pawlak 于 1982 年提出来的. 其主要思想是在保持分类能力不变的前提下，通过知识约简而得到问题的分类规则与决策规则. 具体地说，Pawlak 粗糙集是建立在分类机制的基础上，将分类理解为对一个特定空间的基于等价关系的划分(等价类)，将知识理解为对数据的划分，每一个被划分的集合看成一个概念. 然后利用由特定空间所产生的已知的知识库，将不精确或不确定的知识用已知的数据库中的知识来近似描述. 粗糙集理论与其他处理不精确或不确定问题理论的最明显区别是它无需提供问题所需处理的数据集合之外的任何先验信息，因此对问题的描述和处理是比较客观的，使得粗糙集理论与概率论、模糊数学和证据理论等其他处理不精确或不确定理论有很强的互补性.

20 世纪 90 年代以来，由于计算机与网络技术的迅速发展，各个领域的信息和数据急剧增长，大量数据等待处理，同时由于人类的参与使数据与信息的不确定性更加显著，数据与信息中的关系更加复杂，当时已有的机器学习方法已经不能完全适应所面临的具有多样性的海量数据分析的要求，因此，如何从大量的、杂乱无章的、强干扰的海量数据中挖掘潜在的、新颖的、正确的、有用的知识，给智能信息处理研究者提出了严峻的挑战. 在这种背景下，粗糙集理论因其在机器学习、数据挖掘、知识发现、决策支持与分析、图像处理、专家系统、过程控制、医疗诊断、金融数据分析、近似推理等方面的独特优势，正成为当前计算机科学、人工智能以及信息科学等领域的研究热点之一.

粗糙集理论的研究由于其历史较短，所以迄今为止，粗糙集概念还没有完全统一的定义：有经典的 Pawlak 意义下的，也有由上、下近似构成的一对集合来命名的，还有以上、下近似构成的区间集合来定义的. 不同的定义往往带来研究的侧重面不同. 目前，对粗糙集理论的研究主要集中在以下几个方面：(1) 粗糙集模型的推广；(2) 问题的不确定性研究；(3) 与其他处理不确定性、模糊性问题的数学理论的关系与互补；(4) 粗计算；(5) 纯数学理论方面的研究；(6) 粗糙

集的算法研究与人工智能其他方向关系的研究等. 这些研究有的是受应用的推动而产生的,有的是纯理论的. 本书侧重于不完备信息系统的粗糙集模型拓展、知识约简理论及规则提取算法等. 有关这些问题的讨论我们将会在本书中逐步展开.

1.2 信息系统

粗糙集理论研究的信息系统通常用一个数据表来表示.

定义 1.1 称 $S=(U,A,V,f)$是一个信息系统,或者数据库系统. 其中 U 是非空的对象集,即 $U=\{x_1,x_2,\cdots,x_n\}$,U 中的每个 $x_i(1\leqslant i\leqslant n)$称为一个对象;$A=C\cup D$ 是表示属性的非空有限集合,$C=\{c_1,c_2,\cdots,c_m\}$称为条件属性集合,$D=\{d_1,d_2,\cdots,d_k\}$表示决策属性集合,且 $C\cap D=\varnothing$,常记 $A=\{a_1,a_2,\cdots,a_{\overline{m}}\}$,其中,$\overline{m}=m+k$,$A$ 中的每个 $a_j(1\leqslant j\leqslant\overline{m})$称为一个属性;$f$ 是 U 和 A 的关系集,也称信息函数集,即 $f=\{f_j\mid 1\leqslant j\leqslant\overline{m}\}$,其中 $f_j:U\to V_j(1\leqslant j\leqslant\overline{m})$是信息函数,$V_j$ 是属性 a_j 的值域;$V=\bigcup\limits_{j\leqslant\overline{m}}V_j$ 称为属性值域.

针对信息系统 S,若 $D=\varnothing$,则称信息系统为数据表,否则称决策信息系统或简称决策表. 下文中,在不引起混淆的情况下,一般记$\overline{m}=m$.

在信息系统 S 中,关系集是非常重要的. 譬如 $f_j(x_i)=v_{ij}$(或记 $f(x_i,a_j)=v_{ij}$)就表示对象 x_i 在属性 a_j 下的属性值是 v_{ij},v_{ij} 可以是定性值(fuzzy 数)、实数值、集值或区间值. 若存在一个 $x\in U,c\in C$,$f(x,c)$未知(记作:$f(x,c)=*$),则称信息系统是不完备的,即不完备信息系统(Incomplete Information System, IIS);否则称信息系统是完备的.

在本小节中,如不作特别说明,认为信息系统是完备的.

例 1.1 表 1.1 给出了 10 个玩具对象 $x_1\sim x_{10}$和具有 3 种条件属性:颜色 c(red, yellow, blue)、形状 s(square, circle, triangle)及体积 v(big, small),决策属性——评价结果 d(beautiful, common)所表示的决策信息系统.

表 1.1 玩具决策信息系统

U	c	s	v	d
x_1	yellow	square	big	beautiful
x_2	red	circle	small	common
x_3	yellow	square	big	beautiful
x_4	blue	triangle	small	beautiful

（续表）

U	c	s	v	d
x_5	red	circle	small	common
x_6	red	square	big	common
x_7	blue	triangle	small	beautiful
x_8	yellow	square	big	beautiful
x_9	blue	triangle	small	beautiful
x_{10}	red	square	big	common

在这个信息系统中，条件属性集为 $C=\{c,s,v\}$，决策属性集为 $D=\{d\}$，$V_c=\{\text{red, yellow, blue}\}$，$V_s=\{\text{square, circle, triangle}\}$，$V_v=\{\text{big, small}\}$，从而条件属性值域为 $V_C=\{\text{red, yellow, blue; square, circle, triangle; big, small}\}$；$f=\{f_c,f_s,f_v\}$，$f_c$，$f_s$ 与 f_v 分别表示玩具（对象）关于颜色、形状与体积的取值（信息函数）. 例如：$f_c(x_5)=\text{red}$，$f_s(x_7)=\text{triangle}$，$f_v(x_{10})=\text{big}$.

由例 1.1 易知，一个信息系统对应一个关系数据表；反过来一个关系数据表也对应着一个信息系统. 因此，信息系统 $S=(U,A,V,f)$ 是关系数据表的一种抽象描述.

定义 1.2　设 U 是对象集，令

$$U^2=U\times U=\{(x_i,x_j)\mid x_i,x_j\in U\},$$

则 $R\subseteq U^2$ 称为 U 上的一个等价关系. 若 R 满足以下条件：

（1）自反性：$(x_i,x_i)\in R(1\leqslant i\leqslant n)$；

（2）对称性：$(x_i,x_j)\in R\Rightarrow(x_j,x_i)\in R(\forall i,j\leqslant n)$；

（3）传递性：$(x_i,x_j)\in R,(x_j,x_k)\in R\Rightarrow(x_i,x_k)\in R(\forall i,j,k\leqslant n)$.

设 R 是 U 上的一个等价关系，记

$$[x_i]_R=\{x_j\in U\mid(x_i,x_j)\in R\},$$

则 $[x_i]_R$ 称为包含 x_i 的等价类.

可以证明：若 $(x_i,x_j)\in R$，则 $[x_i]_R=[x_j]_R$，记 $U/R=\{[x_i]_R\mid x_i\in U\}$，则 U/R 称为 U 的划分.

例 1.2　在例 1.1 给出的玩具决策信息系统中，若只考虑各对象在条件属性上的属性值相等，则有等价关系

$R=\{(x_1,x_1),(x_1,x_3),(x_1,x_8),(x_2,x_2),(x_2,x_5),(x_3,x_3),(x_3,x_8),(x_4,x_4),(x_4,x_7),(x_4,x_9),(x_5,x_5),(x_6,x_6),(x_6,x_{10}),(x_7,x_7),(x_7,x_9),(x_8,x_8),(x_9,x_9),(x_{10},x_{10})\}$.

由 R 产生的划分为 $U/R=\{X_1,X_2,X_3,X_4\}$，其中 $X_1=\{x_1,x_3,x_8\}$，$X_2=\{x_2,x_5\}$，$X_3=\{x_4,x_7,x_9\}$，$X_4=\{x_6,x_{10}\}$，均为等价类.

1.3 集合近似与粗糙集

设 U 是有限个对象构成的集合(即论域)，R 是 U 上的等价关系，则 $P=(U,R)$ 称为 Pawlak 近似空间. $X\subseteq U$，X 可能被 R 分成类，也可能不被 R 分成类，凡是能分成类的子集 X 称之为 R 可定义的；否则称之为 R 不可定义的. 也就是说，当 X 能表示成 R 的某些等价类之并时，X 才是 R 可定义的；否则就是 R 不可定义的.

定义 1.3 R 可定义集称为 R 精确集；而 R 不可定义集称为 R 非精确集或 R 粗糙集(rough set).

R 粗糙集虽然不能被 R 的等价类精确地描述，但能否被近似地描述？粗糙集理论对此问题给出了肯定的回答. 即用两个精确集——上近似集(upper-approximation set)和下近似集(lower-approximation set)来近似地描述.

定义 1.4 设 $P=(U,R)$ 是 Pawlak 近似空间，R 是 U 上的等价关系，对于任意 $X\subseteq U$，称

$$\underline{R}(X)=\{x\in U\mid [x]_R\subseteq X\}=\bigcup\{[x]_R\mid [x]_R\subseteq X\}, \tag{1.1}$$

$$\overline{R}(X)=\{x\in U\mid [x]_R\cap X\neq\varnothing\}=\bigcup\{[x]_R\mid [x]_R\cap X\neq\varnothing\}. \tag{1.2}$$

式(1.1,1.2)分别为 X 关于近似空间(U,R)的下近似集与上近似集.

定义 1.4 说明：$\underline{R}(X)$是包含于 X 中的最大 R 精确集；而$\overline{R}(X)$是包含 X 的最小 R 精确集. 集合 $\mathrm{BN}_R(X)=\overline{R}(X)-\underline{R}(X)$，称为 X 的 R 边界(域)，它是 R 精确集；$\mathrm{POS}_R(X)=\underline{R}(X)$，称为 X 的 R 正域；$\mathrm{NEG}_R(X)=U-\overline{R}(X)$，称为 X 的 R 负域，它也是 R 精确集. 显然，$\overline{R}(X)=\mathrm{BN}_R(X)\cup\underline{R}(X)$. 边界域中的元素可能属于 X，也可能不属于 X；正域中的元素必属于 X；而负域中的元素一定不属于 X. 如图 1.1 所示.

粗糙集的定义说明：如果我们用已知的知识(等价类)来描述知识库(论域上的一簇等价类)中的知识 X(论域上的子集)时，一般只能近似地描述(即 X 的上近似集与下近似集)，而很难描述知识 X 的全部，这种描述是粗糙的；仅当 X 边界域为空时，可以通过已知的知识描述 X，这种描述是精确的. 显然，下面关于精确集和粗糙集的性质是成立的.

定理 1.1 设 $P=(U,R)$ 是 Pawlak 近似空间，$X\subseteq U$. 则

(1) X 是 R 精确集，当且仅当$\overline{R}(X)=\underline{R}(X)$；

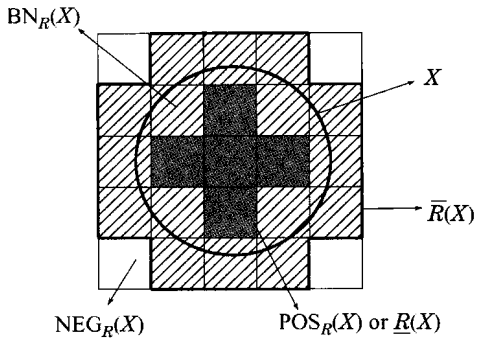

图 1.1　上下近似集、正域、负域和边界域示意图

(2) X 是 R 粗糙集，当且仅当 $\overline{R}(X)\neq\underline{R}(X)$，即 $BN_R(X)\neq\varnothing$.

例 1.3　在例 1.2 中，令 $X=\{x_1,x_2,x_3,x_5,x_6,x_8\}$，则 $\overline{R}(X)=\{x_1,x_2,x_3,x_5,x_6,x_8,x_{10}\}$，$\underline{R}(X)=POS_R(X)=\{x_1,x_2,x_3,x_5,x_8\}$，$BN_R(X)=\{x_6,x_{10}\}$，$NEG_R(X)=\{x_4,x_7,x_9\}$.

由定义 1.4，可得上近似集和下近似集如下性质.

定理 1.2　设 $P=(U,R)$ 是 Pawlak 近似空间，$X,Y\subseteq U$. 则

(1) $\underline{R}(\varnothing)=\overline{R}(\varnothing)=\varnothing$，$\underline{R}(U)=\overline{R}(U)=U$，$\varnothing$ 表示空集；

(2) $\underline{R}(X)\subseteq X\subseteq\overline{R}(X)$；

(3) $\underline{R}(X\cup Y)\supseteq\underline{R}(X)\cup\underline{R}(Y)$，$\overline{R}(X\cup Y)=\overline{R}(X)\cup\overline{R}(Y)$；

(4) $\underline{R}(X\cap Y)=\underline{R}(X)\cap\underline{R}(Y)$，$\overline{R}(X\cap Y)\subseteq\overline{R}(X)\cap\overline{R}(Y)$；

(5) $X\subseteq Y\Rightarrow\underline{R}(X)\subseteq\underline{R}(Y)$，$\overline{R}(X)\subseteq\overline{R}(Y)$；

(6) $\overline{R}(\overline{R}(X))=\underline{R}(\overline{R}(X))=\overline{R}(X)$，$\overline{R}(\underline{R}(X))=\underline{R}(\underline{R}(X))=\underline{R}(X)$；

(7) $\overline{R}(X^C)=(\underline{R}(X))^C$，$\underline{R}(X^C)=(\overline{R}(X))^C$，$X^C$ 表示 X 的余集.

以上性质的证明此处从略.

粗糙集的不可定义性(不确定性)是由于边界的存在而引起的，集合 X 的边界越大，其不确定性也越大，确定性就越小，从而精确性也就越低，粗糙性越大. 为了更准确地描述这一点，我们引入集合 X 的粗糙度与近似精度的概念.

定义 1.5　设 $P=(U,R)$ 是 Pawlak 近似空间，$X\subseteq U$.

(1) 集合 X 的粗糙度定义为

$$\rho_R(X)=|BN_R(X)|/|\overline{R}(X)|=1-|\underline{R}(X)|/|\overline{R}(X)|. \tag{1.3}$$

(2) 集合 X 的近似精度定义为

$$\alpha_R(X)=|\underline{R}(X)|/|\overline{R}(X)|=1-\rho_R(X). \tag{1.4}$$

其中：$X\neq\varnothing$，$|X|$ 表示集合 X 的基数；如果 $X=\varnothing$，定义 $\alpha_R(X)=1$.

显然，$0 \leqslant \alpha_R(X), \rho_R(X) \leqslant 1$.

集合 X 的近似精度与粗糙度是完全相反的两个概念，近似精度表示对知识库中的知识 X 描述的确定程度；而粗糙度表示对知识 X 描述的不确定程度.

根据粗糙集 X 的上近似集、下近似集的特征，对粗糙集的不确定程度也可以给出以下定义.

定义 1.6 设 $P=(U,R)$ 是 Pawlak 近似空间，$X \subseteq U$ 是关于 R 的粗糙集.

(1) 如果 $\underline{R}(X) \neq \varnothing$，且 $\overline{R}(X) \neq U$，则称 X 是 R 粗糙可定义的；

(2) 如果 $\underline{R}(X) = \varnothing$，且 $\overline{R}(X) \neq U$，则称 X 是 R 内不可定义的；

(3) 如果 $\underline{R}(X) \neq \varnothing$，且 $\overline{R}(X) = U$，则称 X 是 R 外不可定义的；

(4) 如果 $\underline{R}(X) = \varnothing$，且 $\overline{R}(X) = U$，则称 X 是 R 全不可定义的.

对定义 1.6，我们可以给出如下的直观解释：

如果 X 是 R 粗糙可定义，则意味着可以确定 U 中哪些元素属于 X 或 X^C（X^C 为 X 的补集）；

如果 X 是 R 内不可定义，则意味着可以确定 U 中哪些元素属于 X^C，但不能确定 U 中的任意元素是否属于 X；

如果 X 是 R 外不可定义，则意味着可以确定 U 中哪些元素属于 X，但不能确定 U 中的任意元素是否属于 X^C；

如果 X 是 R 全不可定义，则意味着不能确定 U 中任意元素属于 X 或 X^C.

定义 1.7 设 $P=(U,R)$ 是 Pawlak 近似空间，$F=\{X_1, X_2, \cdots, X_k\}$ 是 U 的独立于 R 的一个分类，则 F 的近似分类质量定义为

$$\gamma_R(F) = \sum_{i=1}^{k} |\underline{R}(X_i)| / |U|. \tag{1.5}$$

分类质量表示的是应用知识 R 能确切地划入 F 类的对象的百分比.

如果 F 是 U 关于 R 的一个划分，则 $\gamma_R(F)=1$.

在经典的清晰集合中，两个集合相等是指它们所含的元素完全相同. 而粗糙集合与之就有本质的区别，它考虑的是近似相等，即使两个集合在经典集合意义下不相等，但在粗糙集合的意义下有可能相等. 下面给出粗糙集的近似相等关系和包含关系.

定义 1.8 设 $P=(U,R)$ 是 Pawlak 近似空间，$X, Y \subseteq U$.

(1) 若 $\underline{R}(X)=\underline{R}(Y)$，则称集合 X 与 Y 是 R 下粗相等，记为 $X \underline{\approx} Y$；

(2) 若 $\overline{R}(X)=\overline{R}(Y)$，则称集合 X 与 Y 是 R 上粗相等，记为 $X \overline{\approx} Y$；

(3) 若 $X \underline{\approx} Y$，且 $X \overline{\approx} Y$，则称集合 X 与 Y 是 R 粗相等，记为 $X \approx Y$.

显然，下粗相等、上粗相等和粗相等都是 U 上的等价关系.

定义 1.9 设 $P=(U,R)$ 是 Pawlak 近似空间，$X, Y \subseteq U$.