李华雄 周献中 李天瑞
王国胤 苗夺谦 姚一豫 编著

决策粗糙集理论及其研究进展

科学出版社

决策粗糙集理论及其研究进展

李华雄　周献中　李天瑞
王国胤　苗夺谦　姚一豫　编著

科学出版社
北京

内 容 简 介

本书由决策粗糙集研究领域的十多位学者共同编写，力图概括该领域国内外研究的最新成果，为进一步研究发展决策粗糙集理论与应用提供借鉴。本书的内容涉及决策粗糙集的理论与应用两大部分，理论部分包括决策粗糙集的基础理论、决策粗糙集的研究进展、三枝决策粗糙集和决策粗糙集的属性约简；应用部分包括基于决策粗糙集的自动聚类方法、基于决策粗糙集模型的文本分类方法和多用户决策粗糙集模型。最后，本书对决策粗糙集的发展历程和方法论作了概括与展望。

本书可供计算机科学、控制科学与工程、管理科学与工程、应用数学等专业的科技人员、大学高年级学生、研究生以及相关工程技术研究人员阅读参考。

图书在版编目(CIP)数据

决策粗糙集理论及其研究进展/李华雄等编著．—北京：科学出版社，2011

ISBN 978-7-03-032530-3

Ⅰ．决… Ⅱ．李… Ⅲ．人工智能-理论研究 Ⅳ．TP18

中国版本图书馆 CIP 数据核字（2011）第 206617 号

责任编辑：魏英杰 杨向萍 / 责任校对：刘小梅
责任印制：赵 博 / 封面设计：陈 敬

科学出版社出版
北京东黄城根北街 16 号
邮政编码：100717
http://www.sciencep.com

新蕾印刷厂 印刷

科学出版社发行 各地新华书店经销

*

2011 年 11 月第 一 版 开本：B5（720×1000）
2011 年 11 月第一次印刷 印张：12 1/4
字数：231 000

定价：55.00 元

（如有印装质量问题，我社负责调换）

序

在过去20年间，由于网络及其他众多领域数据分析与处理的期望和驱动，统计机器学习得到相关领域研究人员的普遍重视。然而这类研究的基础，大变量集合上的优化建模理论却遇到了根本性的困难，即大变量集合中的变量张成的空间巨大，使得无论多少样本在这个空间上均是稀疏的；另一方面，由于实验设计需要考虑这些变量的组合，使得获得合适的样本集变得不可行。这个困难驱使很多研究者使用越来越复杂的数学工具，而研究结果越来越难以理解，同时，众多领域对数据分析与处理的迫切需求，对统计机器学习的研究形成越来越大的压力，迫使研究者开始寻找新的出路。

新出路的先决条件是保留概率统计学。人们将目光移向概率统计的Bayes方法，这个方法最大的负担是先验分布(概率)。在20世纪80年代末，基于优化的统计机器学习与Bayes图几乎同时提出，但是，研究者与工程师选择了前者，无先验但有泛化的特性应该是人们选择的重要指标。

Bayes决策的历史应该更为久远，在计算机领域，至少Duda在他1973年的著作中就使用这个方法作为模式识别的理论基础。将这个理念与粗糙集理论相结合，形成决策粗糙集理论，则是Y. Y. Yao的贡献。将数据分析与处理的理论续接到Bayes理念，目前，已有多种路径，最著名的应该是概率图模型。

应该说，决策粗糙集理论继承了传统粗糙集理论的上、下近似的概念，而其最重要的差别则是后者重新定义了等价关系。这样，概率统计的理念被植入到样本集合的划分之中，从而既可以使分布描述限制在问题的局部，避免样本稀疏的困难，又可以显现地描述变量之间的统计关系。

决策粗糙集理论的另一个重要特点是关注所谓边界域(边缘域)的决策语义，认为边界域决策也应作为总体分类决策的一部分。事实上，粗糙集最重要的贡献之一就是边界域，尽管很多理论有边界域的成分，但是一般不可能成为研究的课题，如概率统计。究其原因主要是这些理论难以给出这个“域”一个清晰精确的界限，这恰恰是Pawlak的贡献。

该书围绕决策粗糙集理论与方法，汇集了相关学者在决策粗糙集领域的成果，反映了当前决策粗糙集的最新研究进展。该书的出版将促进决策粗糙集理论与应用研究的进一步发展。

王　珏
中国科学院自动化研究所
2011年10月

前　言

决策粗糙集是加拿大华人学者姚一豫(Y. Y. Yao)教授等在20世纪90年代提出的一种粗糙集理论。该理论是对经典Pawlak粗糙集理论的概率拓展，由于引入了Bayes风险决策理论及三枝决策语义，使其在不确定性知识获取和数据处理中具有更加可靠的理论依据和语义解释。近年来，决策粗糙集在粗糙集理论的拓展研究领域以及风险决策、信息过滤、聚类分析、网络支持系统与博弈分析、文本分类等应用研究领域得到了成功运用。自2009年以来，在国际粗糙集与知识技术学术会议(RSKT)、国际认知信息学术会议(ICCI)、中国粗糙集及软计算会议上多次举办了以决策粗糙集为主题的专题讨论会。目前，决策粗糙集理论中包含的误分类容忍机制、三枝决策语义和风险决策分析方法正逐渐引起国内外越来越多学者的关注。

在经典Pawlak粗糙集模型中，近似集的定义以集合的代数包含关系为基础，由于代数包含关系要求较为严格，Pawlak粗糙集方法在容错能力方面具有较大的局限性。众所周知，人类智能对于概念的描述往往是模糊的和不确定的，其对概念的认识具有很强的容错能力与纠错能力，如何在粗糙集中引入对这种容错能力的刻画显得十分重要。因此，人们通过在Pawlak粗糙集中引入概率包含关系提出了一系列概率粗糙集模型，以实现粗糙集对容错能力的刻画。决策粗糙集正是在Pawlak粗糙集向概率粗糙集拓展中产生的。它也是一种概率粗糙集，但与其他概率粗糙集相比，决策粗糙集具有以下几大优势：其一，在决策粗糙集理论框架下，已有概率粗糙集模型能实现统一，并可以视为决策粗糙集在取不同参数情况下的特例。因此，通过研究决策粗糙集可以在更一般层面上探求概率粗糙集模型的本质特性。其二，决策粗糙集借助Bayes风险决策理论，给出了计算概率阈值的方法，这为确定合适的误分类容忍度提供了可靠的理论依据和语义解释。其三，决策粗糙集理论依据正域、负域和边界域的划分方法以及假设检验方法给出了三枝决策模型和语义，可以用来解释实际应用中的许多决策现象，为模拟人类在决策过程中的不确定性模式提供了一种有效手段。

目前，有关决策粗糙集的理论与应用的研究成果日渐增多，参与到该领域的学者和研究队伍也在逐渐扩大。为使更多的学者能够了解决策粗糙集理论和方法并参与研究，共同促进该领域的发展，我们编写了本书。这是本领域研究的第一本专题著作，希望能为决策粗糙集的研究发展作出一定贡献。

本书总结了决策粗糙集理论与应用研究的主要成果，介绍了目前决策粗糙集

研究的最新进展。全书共八章，分为 4 部分内容，章节总体按照从理论到应用的思路进行编排。第 1 部分为决策粗糙集理论的介绍及研究概要，包括第 1 章决策粗糙集理论方法研究综述。第 2 部分为决策粗糙集的理论研究，包括第 2 章三枝决策粗糙集；第 3 章基于决策风险最小化的属性约简；第 4 章决策粗糙集的正域约简。第 3 部分为决策粗糙集的应用研究，包括第 5 章基于决策粗糙集的自动聚类方法；第 6 章基于决策粗糙集模型的文本分类研究；第 7 章多用户决策粗糙集模型。第 4 部分为决策粗糙集研究的总体回顾与方法论总结，包括第 8 章决策粗糙集研究探讨。各章之间内容相对独立，但都紧密围绕决策粗糙集的主题展开。本书的内容引用了前期的一些研究成果，同时也包含了作者部分最新的研究成果，因此本书既是对前期研究内容的总结，也是对未来研究的展望，可为读者们进一步研究决策粗糙集提供参考。

本书是国内外决策粗糙集研究领域学者共同努力的结果，感谢每一位为本书编写和出版作出努力的人。同时感谢国家自然科学基金(60971062，60873108，60971063)的资助。

由于作者自身知识水平有限，书中不妥之处在所难免，恳请广大读者批评指正。

编　者

2011 年 10 月

目　录

第 1 章　决策粗糙集理论方法研究综述

人类智能在决策时通常表现出不确定性、非精确性、容错性与模糊性等特点，如何通过计算机模拟人类智能的这些特点一直是智能科学领域关注的重要问题。近几十年来，模糊集、粗糙集、证据理论等描述不确定性的数学工具被运用到智能推理研究中，推动了不确定性人工智能的迅速发展。决策粗糙集(decision-theoretic rough sets)作为粗糙集理论的重要组成部分，其给出的三枝决策语义和概念容错分析方法有效地模拟了人类智能的不确定性和非精确的特点。它将传统的正域、负域二分决策语义拓展为正域、边界域和负域的三枝决策语义，认为边界域决策也是一类可行的决策。这与人类智能在处理决策问题的方法是一致的。为刻画概念的非精确性及分类的容错性特点，决策粗糙集在概念上、下近似集中引入概率包含关系，并依据最小化风险原则给出了概率阈值的确定方法，为选取最优分类决策结果提供了理论依据，这也为模拟人类智能在选择最优决策时的推理机制提供了一种途径。本章主要介绍决策粗糙集的基本思想与概念，对决策粗糙集产生、发展、理论基础、应用研究等作综述性的回顾，以使读者对决策粗糙集的理论和方法有整体了解。

1.1　引　　言

决策粗糙集理论是由加拿大的 Yao 等在 20 世纪 90 年代初提出的一种粗糙集理论与方法。该理论在粗糙集中引入了概率包含关系，并通过 Bayes 风险决策方法确定概念边界，建立了具有噪声容忍机制的决策粗糙集模型。Yao 在 1990 年最早提出了决策粗糙集理论[1]，随后其在粗糙集理论拓展研究[2~23]以及信息过滤[24~27]、风险决策分析[28~32]、聚类分析与文本分类[33~39]、网络支持系统与博弈分析[40~42]等领域得到了成功运用，正逐渐引起国内外越来越多学者的关注。近年来，在国际国内粗糙集学术会议与有关期刊上，关于决策粗糙集的研究成果日渐增多。2009 年国际粗糙集与知识技术学术会议(RSKT)举办了以决策粗糙集为主题的特别分会，2010 年 RSKT 会议继续举行了该主题分会，同年国际认知信息学术会议(ICCI)和 CRSSC-CWI-CGrC'10 也有以决策粗糙集为主题的特别专题会议。2011 年在加拿大召开的 RSKT'2011 也设有以决策粗糙集为主题的特别

本章执笔人：李华雄、周献中，南京大学工程管理学院。

分会。

在经典粗糙集理论研究中，Pawlak 代数粗糙集模型是研究的主要对象[43,44]，其核心基础是基于等价关系的已知概念粒化，以及上下近似集对未知概念的逼近。其中，基于等价关系的已知概念粒化是知识表述的一种基本模型，它将知识表示为对论域的划分，即根据对象的不同属性将其划分为不同的子集，从而形成已知概念。对于未知概念，需通过已知概念对其进行近似刻画，在粗糙集理论中，这个过程就是上下近似集对未知概念的逼近。当已知概念的粒度充分细时，它对未知概念的刻画越精准；反之，则刻画越粗略。Pawlak 代数粗糙集模型模拟了人类智能中的概念粒化能力和概念近似能力，而概念粒之间的代数包含关系是这种模拟的理论基础。

然而，概念粒之间代数包含关系导出的近似集在模拟人类智能的容错能力方面具有明显不足。人类智能对于概念的描述往往是模糊和不确定的，其对概念的认识具有很强的容错与纠错能力，这种能力难以用精确的代数包含关系进行刻画[3]。例如，有经验的医生能够根据患者“发热”和“咳嗽”的症状快速诊断其可能患上肺炎，尽管具有这两种症状的患者并非全患上肺炎。医生在分析症状与病情关系时考虑的是概念间的概率特性包含关系，因此具有较强的容错能力。在经典 Pawlak 代数粗糙集模型[43,44]中，由于正域是建立在代数包含关系基础上的，因此难以体现概念表示的容错性，这正是经典 Pawlak 代数粗糙集模型的局限所在。

针对 Pawlak 代数粗糙集模型缺乏容错能力的问题，人们考虑在正域中引入误分类容忍机制，并在此基础上提出了参数可调的概率粗糙集模型。在这些模型研究中，主要代表性的成果有：Yao 等提出了基于 Bayes 风险分析的决策粗糙集(decision-theoretic rough sets，DTRS)模型[1~11]，Wong 和 Ziarko 等对概率近似分类与模糊集作了比较研究[45]，Pawlak、Wong 和 Ziarko 等提出了 0.5-概率粗糙集模型[46]，Ziarko 提出了可变精度粗糙集模型[47]，Pawlak-和 Skowron 等引入了粗糙隶属函数的概念[48]，Skowron、Pawlak 和 Stepaniuk 等提出了参数化粗糙集模型[49,50]，Slezak 等研究了 Bayes 粗糙集模型[51~53]。

作为最早的概率粗糙集模型之一，决策粗糙集的基础工作主要由 Yao 等完成。他们发现[1,2]：在经典 Pawlak 代数粗糙集模型中，论域总是表示为总体决策类的正域和边界域这两者的并集，而总体决策类的负域恒为空集，这与单个决策类可由正域、边界域和负域三者完备刻画是不相称的。这种不完备刻画产生的原因是由于在经典 Pawlak 代数粗糙集定义下，单个决策类的负域必定可以划分到它的补决策类的正域或边界域中，从而造成总体决策类负域为空。Yao 认为论域的这种不完备粗糙结构反映了 Pawlak 代数粗糙集模型对总体决策类刻画具有不完整性，无法刻画总体决策类的负域。为了解决该问题，Yao 将 Pawlak 代数粗糙集模型中的代数包含关系拓展为可调的概率包含关系，使得单个决策类的正域、边界

域和负域由相应的概率包含范围确定。这样单个决策类的负域不一定划分到互补决策类的正域或边界域中，因此总体决策类的负域可以为非空，从而整个论域可表示为决策类的正域、边界域和负域三者的并集。这样的全域粗糙结构实现了总体决策类刻画的完备性[1,2]。

在实现总体决策类完备性刻画的同时，决策粗糙集通过包含度阈值引入了误分类容忍机制，允许概念刻画中存在一定程度的误差：只需要等价类的大部分被目标概念所包含，即将其分类为正域，而当等价类的大部分不在目标概念中则将其分类为负域，介于两者之间的则分类为边界域。这种通过概率包含关系区分正域、边界域和负域的方法是概率粗糙集模型的典型特征。与其他概率型粗糙集理论不同的是，决策粗糙集模型引入了 Bayes 风险理论，其区分正域、边界域和负域的包含度阈值通过计算各分类决策的最小风险代价得到，是具有最小决策风险的优化结果。由于将 Bayes 风险理论引入到阈值的确定中，决策粗糙集在应用问题中具有充分的理论依据。Yao 等提出决策粗糙集模型后，进一步分析了决策粗糙集同模糊集、经典粗糙集的关系，并将决策粗糙集模型与其他概率粗糙集模型进行了比较，指出模糊集意义下的 α- 截集、经典 Pawlak 粗糙集及其他类型的概率粗糙集模型可统一到决策粗糙集模型中，它们可以视为决策粗糙集模型在不同参数条件下的特例[1,2,4]。

在此基础上，Yao 和 Zhao 研究了决策粗糙集意义下的约简理论，指出 Pawlak 代数粗糙集模型的约简理论在概率粗糙集模型中不再适用，由此提出了决策粗糙集约简所需维持不变的若干特征，并系统阐述了决策粗糙集约简理论[6]。Li、Zhou 等通过研究决策粗糙集正域的非单调性特征，提出了一种非单调决策粗糙集正域约简定义与搜索算法[14]。Jia 等研究了决策风险最小化情形下的属性约简和基于三枝决策的属性约简[20~22]。Yao、Liu 和 Li 等研究了决策粗糙集的三枝决策语义，并提出了三枝决策粗糙集模型[7,8,10,19,30]。Liu 等将两类决策粗糙集模型拓展为多类决策粗糙集模型[16]。在应用方面，Zhou 和 Li 提出了基于决策粗糙集的风险偏好决策模型，给出了不同决策主体在乐观倾向、悲观倾向和中性倾向下的决策模式[28, 29]。Li 和 Miao 等提出了基于决策粗糙集模型的文本分类方法[33]。Yu 和 Lingras 等分别研究了决策粗糙集在聚类问题中的应用[36~39]。Yao 和 Herbert 等提出了基于博弈论的决策粗糙集数据分析方法[41,42]，研究了决策粗糙集在属性选择问题中的应用[54]，并给出了决策粗糙集在网络支持系统中的应用方法[40]。Zhao、Li 和 Zhou 等分别研究了决策粗糙集理论在信息过滤中的应用[24~27]。Yao 和 Zhou 提出了朴素 Bayes 粗糙分类模型[9]。Yang 等基于多用户决策理论提出了多用户决策粗糙集模型[31]，并研究了不完备信息系统的决策粗糙集模型[32]。Ayad等研究了 Bayes 决策粗糙集在 E-learning 系统中的应用[55]。

本章余下部分首先介绍决策粗糙集的 Bayes 决策理论，决策粗糙集的基本概

念，分析概率粗糙集模型、决策粗糙集模型与经典 Pawlak 代数粗糙集模型以及一般概率粗糙集模型之间的关系等。然后讨论决策粗糙集意义下的三枝决策语义以及约简定义，并回顾决策粗糙集在实际问题中的应用。

1.2 决策粗糙集理论

为叙述完整，首先回顾粗糙集理论中的 Pawlak 代数粗糙集模型[43,44]，在此基础上引入 Bayes 风险决策理论及决策粗糙集模型。

1.2.1 Pawlak 代数粗糙集模型

粗糙集理论的中心内容是上下近似集的定义。在 Pawlak 代数粗糙集模型中，上下近似集通过等价类与目标概念类的两种代数包含关系给出定义，在此基础上进一步定义了正域、负域与边界域。设 U 表示论域，其为一有限非空集合，$R \subseteq U \times U$ 为论域 U 上的等价关系，整个论域可以通过该等价关系划分成互不相交的子集，即可形成论域 U 上的一个划分 U/R。如果论域中的两个对象 x 和 y 在同一个等价类中，则 x 和 y 是不可区分的。在一个信息表中，设 C 为属性集，则等价关系 R 可由信息表中的属性子集 $A \subseteq C$ 导出。设 X 为论域 U 的子集，即 $X \subseteq U$，则 X 的下近似集 $\underline{\mathrm{apr}}_R(X)$ 与上近似集 $\overline{\mathrm{apr}}_R(X)$ 分别为

$$\begin{aligned} \underline{\mathrm{apr}}_R(X) &= \{x \mid x \in U, [x]_R \subseteq X\} \\ \overline{\mathrm{apr}}_R(X) &= \{x \mid x \in U, [x]_R \cap X \neq \varnothing\} \end{aligned} \tag{1.1}$$

根据 X 的上下近似集定义，整个论域 U 可以划分为互不相交的正域 $\mathrm{POS}(X)$、负域 $\mathrm{NEG}(X)$ 和边界域 $\mathrm{BND}(X)$ 三部分，分别定义为

$$\begin{aligned} \mathrm{POS}(X) &= \underline{\mathrm{apr}}_R(X) \\ \mathrm{NEG}(X) &= U - \overline{\mathrm{apr}}_R(X) \\ \mathrm{BND}(X) &= \overline{\mathrm{apr}}_R(X) - \underline{\mathrm{apr}}_R(X) \end{aligned} \tag{1.2}$$

对于正域中的每个对象 x，其等价类完全包含于 X。因此，关于 x 的特征描述必定属于概念 X，由此可以导出确定性规则，而在边界域中的每个对象 x，其是否属于概念 X 具有不确定性。根据式(1.1)和式(1.2)，可以得到上近似集与正域、边界域的关系：$\overline{\mathrm{apr}}_R(X) = \mathrm{POS}(X) \cup \mathrm{BND}(X)$，即上近似集包含确定与可能属于概念 X 的对象，因此对于 $x \in \overline{\mathrm{apr}}_R(X)$，其可能属于概念 X，由此导出所有可能性规则。此外，对于 X 的负域中的每个对象 x，其等价类完全包含于 $\neg X$，即关于 x 的特征描述必定属于概念 $\neg X$，因此由 X 的负域也可导出确定性规则。

从 Pawlak 代数粗糙集模型中正域、负域与边界域的定义可以看出，目标概念 X 三种区域的界定实质上确定了论域中所有对象的一种分类决策，即根据 x 的特

征是否将 x 分类到 X 的决策。这种对 x 的分类决策对应于一个分类器：若 x 属于 X 的正域，则 x 确定分类为 X；若 x 属于 X 的负域，则 x 确定分类为 $\neg X$；若 x 属于 X 的边界域，则 x 是否分类为 X 是不明确的。从机器学习的角度来看，分类器的一种理想性能是误分类率为零，而 Pawlak 代数粗糙集模型中的正域可以保证其中对象分类为目标概念的误分类率为零，从这个意义上说正域确定了一个误分类率为零的精确分类器。然而，在实际问题中，由于噪声的存在，要得到误分类率为零的精确分类器通常难以做到，因此将正域定义为完全包含于目标概念的等价类显得过于严格，它缺乏对误分类的容忍能力。因此，有必要拓展 Pawlak 代数粗糙集模型中的上下近似集定义，提出具有容错能力的粗糙集模型。

另一方面，分类器的误分类率是相对于论域中已知样本（训练样本）而言的，误分类率为零仅能保证经验风险最小化，却不能保证结构风险最小化。单纯追求分类器在已知样本的误分类率最小将导致分类器的泛化能力降低，从而在新样本的分类中具有较大的误分类率。因此，从机器学习的角度来看，粗糙集中正域的定义不应以误分类率低作为唯一目标。在实际问题中有时需要考虑比误分类率更为广泛的概念——风险。例如，我们在根据“发热”和“咳嗽”症状判断是否患上肺炎时，不但要考虑到尽可能作出正确的判断，而且还要考虑到作出错误判断时会带来什么后果。如果将没有患上肺炎的患者误判为患上肺炎固然会给患者带来一定程度的精神负担，然而如果患者真的已患上肺炎却将其误判为正常，就会使早期肺炎患者失去进一步检查治疗的机会，造成严重的后果。显然，这两种不同的错误判断造成损失的严重程度是有明显差异的，后者的损失显然比前者要更严重。因此，当患者患肺炎的可能性超过一定程度时，尽管没有百分之百的把握判断一定患上肺炎，人们更多倾向于将其判断为已患上肺炎，从而减少错误判断的整体风险。因此，在界定目标概念的正域时，允许多大程度的误分类率需要从最小化整体风险来考虑。最小风险 Bayes 决策正是基于这种思想的决策方法，它也是决策粗糙集模型的理论基础。为此，下面我们将介绍基于最小风险的 Bayes 决策。

1.2.2　基于最小风险的 Bayes 决策

下面从决策论角度介绍基于最小风险的 Bayes 决策[56]。在决策论中，决策对象所有可能的状态 $\omega_j(j=1,2,\cdots,s)$ 构成状态空间 $\Omega=\{\omega_1,\omega_2,\cdots,\omega_s\}$，所有可能的各种决策组成的集合称为决策空间，用 $A=\{a_1,a_2,\cdots,a_m\}$ 表示，其中 $a_i(i=1,2,\cdots,m)$ 表示各种决策。每一个决策行为 a_i 都会带来一定的损失，其大小取决于选择的决策以及决策对象的状态，因此损失可以视为决策与状态的函数。这种函数关系可以通过损失矩阵进行表述。例如，以肺炎诊断决策为例，其损失函数可由表 1.1 给出的损失矩阵表述。

表 1.1 肺炎诊断损失矩阵

状态 / 损失 / 决策	患肺炎	正常
患肺炎	0	10
正常	1000	0

在表 1.1 中，作出正确分类决策的损失最小，均为 0，作出错误分类决策的损失大于 0，但两种错误分类决策的损失大小不同，反映了人们对两种误分类决策的不同风险偏好。一般而言，当研究的对象具有 s 个不同状态，同时有 m 个决策行动可以选择时，决策损失矩阵可以用 $m \times s$（或 $s \times m$）的表格描述，如表 1.2 所示。其中损失函数为 $\lambda(a_i \mid \omega_j)$，表示当真实状态为 ω_j 采取决策为 a_i 时所带来的损失。

表 1.2 决策损失矩阵的一般形式

状态 / 损失 / 决策	ω_1	ω_2	…	ω_j	…	ω_s
a_1	$\lambda(a_1 \mid \omega_1)$	$\lambda(a_1 \mid \omega_2)$	…	$\lambda(a_1 \mid \omega_j)$	…	$\lambda(a_1 \mid \omega_s)$
a_2	$\lambda(a_2 \mid \omega_1)$	$\lambda(a_2 \mid \omega_2)$	…	$\lambda(a_2 \mid \omega_j)$	…	$\lambda(a_2 \mid \omega_s)$
⋮	⋮	⋮		⋮		⋮
a_i	$\lambda(a_i \mid \omega_1)$	$\lambda(a_i \mid \omega_2)$	…	$\lambda(a_i \mid \omega_j)$	…	$\lambda(a_i \mid \omega_s)$
⋮	⋮	⋮		⋮		⋮
a_m	$\lambda(a_m \mid \omega_1)$	$\lambda(a_m \mid \omega_2)$	…	$\lambda(a_m \mid \omega_j)$	…	$\lambda(a_m \mid \omega_s)$

在实际问题中，对象所处的状态通常与其显示的特征相关联。例如，患者是否处于患肺炎状态可以通过患者发热、咳嗽、白细胞数等症状特征加以判断，这些可以观测的特征可用向量 $\boldsymbol{x}$ 来表示。假定各个状态发生的先验概率 $P(\omega_j)$ 和状态 ω_j 下具有特征 $\boldsymbol{x}$ 的条件概率 $P(\boldsymbol{x} \mid \omega_j)$ 均已知，根据 Bayes 概率公式，则已知特征 $\boldsymbol{x}$ 的条件下对象具有状态 ω_j 的后验概率 $P(\omega_j \mid \boldsymbol{x})$ 为

$$P(\omega_j \mid \boldsymbol{x}) = \frac{P(\boldsymbol{x} \mid \omega_j)P(\omega_j)}{P(\boldsymbol{x})} \tag{1.3}$$

其中，$P(\boldsymbol{x}) = \sum_{j=1}^{s} P(\boldsymbol{x} \mid \omega_j)P(\omega_j)$，则有

$$P(\omega_j \mid \boldsymbol{x}) = \frac{P(\boldsymbol{x} \mid \omega_j)P(\omega_j)}{\sum_{i=1}^{s} P(\boldsymbol{x} \mid \omega_i)P(\omega_i)}$$

由于需考虑决策带来的总体损失，因此不能仅根据后验概率的大小进行决策，而应考虑决策行动是否使总体的损失最小，以此作为选择最优决策的依据。对于

给定的观测特征 $\boldsymbol{x}$，采取决策 a_i 所对应的损失值为各种可能状态下损失的期望值，即

$$R(a_i \mid \boldsymbol{x}) = E[\lambda(a_i \mid \omega_j)] = \sum_{j=1}^{s} \lambda(a_i \mid \omega_j) P(\omega_j \mid \boldsymbol{x})$$

在考虑误分类带来的损失时，我们希望最优决策能够具有最小的期望损失值。如果在作出每个决策时，都使得其条件风险值最小，则对所有的 $\boldsymbol{x}$ 进行决策时，其期望风险也必然最小。这样的决策便是最小风险 Bayes 决策。据此，最小风险 Bayes 决策规则可形式化为

$$\text{If } R(a_k \mid \boldsymbol{x}) = \min_{i \in \{1,2,\cdots,m\}} R(a_i \mid \boldsymbol{x}), \text{ Then } a = a_k \tag{1.4}$$

当分类为互补的两类问题时，状态集为 $\Omega = \{X, \neg X\}$，分别表示状态为正例 X 和负例 $\neg X$；决策集为 $A = \{a_P, a_N\}$，分别表示将对象分类为正例 X 和负例 $\neg X$ 的决策。此时，最小风险 Bayes 决策规则可表示为

$$\begin{aligned} &\text{If } R(a_P \mid \boldsymbol{x}) \leqslant R(a_N \mid \boldsymbol{x}), \text{ Then } a = a_P \\ &\text{If } R(a_N \mid \boldsymbol{x}) \leqslant R(a_P \mid \boldsymbol{x}), \text{ Then } a = a_N \end{aligned} \tag{1.5}$$

根据最小风险 Bayes 决策规则的定义，在求解最小风险 Bayes 决策时可以按照以下步骤来实现：

① 给定特征描述 $\boldsymbol{x}$，根据先验概率 $P(\omega_j)$ 和状态 ω_i 下具有特征 $\boldsymbol{x}$ 的条件概率 $P(\boldsymbol{x} \mid \omega_j)$ 计算后验概率，即

$$P(\omega_j \mid \boldsymbol{x}) = \frac{P(\boldsymbol{x} \mid \omega_j) P(\omega_j)}{\sum_{i=1}^{s} P(\boldsymbol{x} \mid \omega_i) P(\omega_i)}, \quad j = 1,2,\cdots,s$$

② 根据后验概率和损失矩阵计算各个决策 a_i 的期望风险 $R(a_i \mid \boldsymbol{x})$，即

$$R(a_i \mid \boldsymbol{x}) = \sum_{j=1}^{s} \lambda(a_i \mid \omega_j) P(\omega_j \mid \boldsymbol{x})$$

③ 比较各个决策 a_i 的期望风险 $R(a_i \mid \boldsymbol{x})$ 的大小，找出是期望风险最小的决策 $R(a_k \mid \boldsymbol{x}) = \min_{i \in \{1,2,\cdots,m\}} R(a_i \mid \boldsymbol{x})$，即为最小风险 Bayes 决策。

1.2.3　决策粗糙集模型

如前所述，在 Pawlak 代数粗糙集中，只有那些能完全包含于目标概念的等价类才能被判断为确定属于决策概念，而对于包含度介于 0 与 1 之间的对象集合均被划归于边界域。这在存在噪声的实际问题中显得过于严格，缺乏对误分类的容错能力。因此，采用定量的概率包含关系来度量对象集合相对于目标概念的隶属度（条件概率）是很有必要的。基于这种概率思想，人们提出了一系列概率粗糙集模型，其中决策粗糙集模型是较早提出来的一类概率粗糙集模型。Yao 论述了各种概率粗糙集模型可视为决策粗糙集的特例，概率粗糙集模型的若干概念和性质