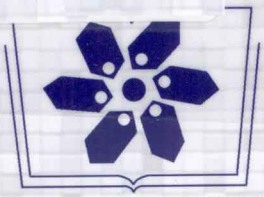


中国科学院科学出版基金资助出版

信息科学技术学术著作丛书

基于不确定性的决策树归纳

王熙照　翟俊海　著

科学出版社

信息科学技术学术著作丛书

基于不确定性的决策树归纳

王熙照　翟俊海　著

科学出版社

北京

内 容 简 介

本书主要介绍不确定性及不确定环境下的决策树归纳方法，包括模糊决策树归纳、最优割点的模糊化处理、决策树优化、主动学习与特征选择在模糊决策树中的应用、模糊决策树的集成学习等内容。本书结合作者近年来关于决策树归纳学习的研究成果，以决策树归纳学习的基本理论为基础，全面系统地讨论了决策树归纳学习中的主要问题。

本书可作为应用数学、智能科学与技术、自动化等专业高年级本科生和研究生的教材，也可供从事相关研究工作的科研人员参考。

图书在版编目(CIP)数据

基于不确定性的决策树归纳/王熙照，翟俊海著．—北京：科学出版社，2012
（信息科学技术学术著作丛书）
ISBN 978-7-03-034635-3

Ⅰ．基… Ⅱ．①王…②翟… Ⅲ．决策树-归纳 Ⅳ．C934

中国版本图书馆 CIP 数据核字(2012)第 117452 号

责任编辑：魏英杰 雷 旸/ 责任校对：林青梅
责任印制：张 倩 / 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

源海印刷有限责任公司印刷

科学出版社发行 各地新华书店经销
*
2012 年 6 月第 一 版 开本：720×1000 1/16
2012 年 6 月第一次印刷 印张：21 3/4
字数：424 000

定价：60.00 元

（如有印装质量问题，我社负责调换）

《信息科学技术学术著作丛书》序

21世纪是信息科学技术发生深刻变革的时代，一场以网络科学、高性能计算和仿真、智能科学、计算思维为特征的信息科学革命正在兴起。信息科学技术正在逐步融入各个应用领域并与生物、纳米、认知等交织在一起，悄然改变着我们的生活方式。信息科学技术已经成为人类社会进步过程中发展最快、交叉渗透性最强、应用面最广的关键技术。

如何进一步推动我国信息科学技术的研究与发展；如何将信息技术发展的新理论、新方法与研究成果转化为社会发展的新动力；如何抓住信息技术深刻发展变革的机遇，提升我国自主创新和可持续发展的能力？这些问题的解答都离不开我国科技工作者和工程技术人员的求索和艰辛付出。为这些科技工作者和工程技术人员提供一个良好的出版环境和平台，将这些科技成就迅速转化为智力成果，将对我国信息科学技术的发展起到重要的推动作用。

《信息科学技术学术著作丛书》是科学出版社在广泛征求专家意见的基础上，经过长期考察、反复论证之后组织出版的。这套丛书旨在传播网络科学和未来网络技术，微电子、光电子和量子信息技术，超级计算机、软件和信息存储技术，数据知识化和基于知识处理的未来信息服务业，低成本信息化和用信息技术提升传统产业，智能与认知科学、生物信息学、社会信息学等前沿交叉科学，信息科学基础理论，信息安全等几个未来信息科学技术重点发展领域的优秀科研成果。丛书力争起点高、内容新、导向性强，具有一定的原创性；体现出科学出版社“高层次、高质量、高水平”的特色和“严肃、严密、严格”的优良作风。

希望这套丛书的出版，能为我国信息科学技术的发展、创新和突破带来一些启迪和帮助。同时，欢迎广大读者提出好的建议，以促进和完善丛书的出版工作。

中国科学院计算技术研究所所长

前　言

机器学习是计算机系统获取智能的本质途径，是人工智能的一个研究领域，主要研究如何让机器具有学习能力。机器学习是一个基于特定目的的知识获取过程，其内在行为是获取知识、发现规律，外在表现是改进性能、适应环境。

归纳学习（或称有监督学习）是机器学习领域的重要分支之一。它的主要任务是从数据中抽取规则，被认为是知识发现的一种重要手段。归纳学习是一种以归纳推理为基础的学习，是一种从多个示例中归纳出一般概念或一般性规律的学习方式，被公认为是专家系统发展的瓶颈。关于归纳学习研究，从近 20 年来出现的浩瀚文献中可以发现大量的归纳学习新方法和新技术，并且成功应用于故障诊断、模式识别、生物信息处理等实用领域。

决策树归纳是归纳学习中一种高效实用的学习方法。决策树方法最早产生于 20 世纪 60 年代中期，Quinlan 提出的 ID3 算法是决策树算法的典型代表，它以信息增益作为选择扩展属性的标准。由于决策树算法结构简单，计算量小，且适用于大规模数据集学习问题，故而已经成为归纳学习的一个重要分支，并在众多的实际领域得到应用。

以 ID3 算法为代表，早期出现的众多构造决策树的算法都是在假定属性取值及分类值是明确的前提下建立的，这些算法都不能处理与人的思维和感觉有关的不确定性。正如 Quinlan 指出的，“决策树的分类结果是分明的，它不能处理分类过程中潜在的不确定性。当属性的取值有微小变化时，有可能导致分类结果明显不合适的突变。生成的决策树一般不具有稳健性，数据信息的不精确或缺少可能完全阻止了样例的分类”。为克服这些缺陷，Quinlan 曾建议采用一种概率方法来构造决策树以处理不确定性。在这个模型下，属性值的不精确性被认为是一种噪声，分支的阈值被软化，最后的分类结果被指定为一种概率估计。这种关于分类问题的不确定性是统计上的，主要源于随机性误差。

有许多种关于不确定性的定义，但总体上可分为两大类，即统计上的和认识上的。统计上的不确定性所处理的是由随机性或系统误差等所产生的现象。与统计上的不确定性不同，认识上的不确定性所表示的现象来源于人的思维、推理、认识和感觉过程，这种认识上的不确定性可进一步细分为不可指定性和模糊性。一般来说，不可指定性代表一种一对多的关系，即从一个具有多个可选择项的问题中选择一个所具有的不可指定性；模糊性主要代表一种边界不分明现象，也就是与不能进行精确区分有关的不确定性。

此为试读，需要完整PDF请访问：www.ertongbook.com

本书首先介绍什么是不确定性，以及几种常见的不确定性：随机性、模糊性、不可指定性和粗糙性。通过讨论这几种不确定性之间的关系，为后面基于不确定性的决策树归纳学习提供基础。其次介绍不确定环境下决策树归纳过程中不确定性的表示、度量及应用。最后介绍不确定环境下的决策树生成算法、匹配策略、决策树优化算法、特征选取和样本选取此外，本书还介绍了不确定环境下的决策树集成和其他的归纳学习方法。

本书的特点是结合作者近年来关于决策树归纳学习的研究成果，以决策树归纳的基本理论（不确定性、模糊决策树的产生机制）为基础，全面讨论决策树归纳学习中的主要问题（不确定环境下决策树扩展属性启发式标准的设计、决策树优化、特征选择、决策树的集成学习等）。本书大部分内容取材于作者王熙照的博士论文《模糊示例学习研究》(1998)和1998年至今作者及其研究团队在此领域的相关研究成果。参加本书撰写和讨论的有翟俊海、冯慧敏、高相辉、陈爱霞、孟庆武、何玉林和董令彩，最后由王熙照和翟俊海定稿。本书的出版得到了2010年度中国科学院科学出版基金项目、国家自然科学基金项目(61170040)、河北省自然科学基金项目(F2008000635，F2010000323)和河北省应用基础重点研究项目(08963522D)的资助，在此一并表示感谢！另外，感谢科学出版社魏英杰老师的帮助。

由于作者水平所限，书中不足之处在所难免，敬请各位读者指正。

王熙照　翟俊海

2011年7月于河北大学

目　录

第1章 不确定性

现实世界中存在许多不确定性现象，例如描述身高的“中等个头”，描述温度的“大约37摄氏度”，描述年龄的“青年”等。不确定性(uncertainty)在人们生活中几乎无处不在，因此研究不确定性的表现、刻画及度量是很有意义的。不确定性在不同学科有不同的含义，因此很难给出不确定性的明确定义。

不确定性有多种，大致可以划分为两大类：客观不确定性和认知不确定性[1]。客观不确定性的大小不以人的主观意志而改变，通常包括随机性(randomness)和粗糙性(roughness)。随机性是由客观系统产生的不确定性；粗糙性是由当前掌握的知识不足而造成的概念刻画上的不确定性。认知不确定性是人类在感知、思考和推理过程中产生的不确定性。认知不确定性通常包括模糊性(fuzziness 或 vagueness)和不可指定性(non-specificity 或 ambiguity)。模糊性是由于人们无法给出清晰准确的界限而产生的不确定性；不可指定性是人们在处理一对多关系时产生的不确定性。

总的来说这两类不确定性一类是客观的，另一类是主观的。本章重点介绍随机性、模糊性、不可指定性和粗糙性。

1.1 随 机 性

1.1.1 随机现象

随机现象从表面上看杂乱无章、没有规律，但实践证明，如果同类的随机现象大量重复出现，总体上就会呈现一定的规律性[2]。比如掷一枚质地均匀的硬币，每一次投掷前很难预料是哪一面朝上，但是如果多次重复地投掷，就会发现正面朝上和反面朝上的次数大体相同。

1.1.2 概率分布

概率论以随机变量为工具研究随机现象，这里不做赘述，仅简单介绍本书使用较多的离散型随机变量及其概率分布。在概率论中，对具有下列特征的实验称为随机实验：

(1) 可以在相同的条件下重复进行。

(2) 每次实验的可能结果不止一个，并且能事先明确实验的所有可能结果。

(3) 进行一次实验之前不能确定哪一个结果会出现。

定义 1.1.1 随机实验的所有可能结果组成的集合称为样本空间，记为 $S=\{e\}$。设 $X=X(e)$是定义在样本空间 S 上的实值单值函数，称 $X=X(e)$为随机变量。有些随机变量全部可能取到的不同值是有限个或可列无限个，这种随机变量称为离散型随机变量。

定义 1.1.2 设离散型随机变量 X 所有可能的值为 $x_k(k=1,2,\cdots)$，X 取各个可能值的概率，即事件 $X=x_k$ 的概率，为

$$P\{X=x_k\}=p_k,\quad k=1,2,\cdots \tag{1.1}$$

p_k 满足如下两个条件：

(1) $p_k\geqslant 0, k=1,2,\cdots$。

(2) $\sum_{k=1}^{\infty} p_k = 1$。

称式(1.1)为离散型随机变量 X 的分布律。当 k 的取值为有限 n 时，离散型随机变量 X 的分布简记为 $p=\{p_1,p_2,\cdots,p_n\}$。有时把随机变量也省略了，直接说有一个概率分布 $p=\{p_1,p_2,\cdots,p_n\}$。

1.1.3 信息熵

信息熵(entropy)由信息论奠基人克劳德·香农在 1948 年提出[3]。在信息论中，信息熵越大，不确定性越大；反之，信息熵越小，不确定性越小。信息熵被广泛应用于机器学习的很多领域，如决策树归纳学习、主动学习、基于近邻的分类等。

定义 1.1.3 对一个概率分布 $p=\{p_1,p_2,\cdots,p_n\}$，其信息熵定义为

$$E(p)=-\sum_{i=1}^{n} p_i\log_2 p_i$$

特殊地，当 $p_i=0(1\leqslant i\leqslant n)$时，令 $p_i\log_2 p_i=0$。信息熵表示概率分布 p 的随机不确定性的大小，也就是概率分布 p 所蕴涵的信息量的大小。

例如，对于一个二维概率分布 $p=\{p_1,p_2\}$，该概率分布的信息熵为

$$E(p)=-p_1\log_2 p_1-p_2\log_2 p_2$$

由于 $p_2=1-p_1$，上式可以等价变形为

$$E(p)=-p_1\log_2 p_1-(1-p_1)\log_2(1-p_1)$$

显然，当 $p_1=0$ 时，信息熵取最小值 0；随着 p_1 从 0 逐渐增大到 0.5，信息熵不断增大；当 $p_1=0.5$ 时，即 $p=\{0.5,0.5\}$，信息熵达到最大值 1；随着 p_1 从 0.5 增大到 1，信息熵不断变小；当 $p_1=1$ 时，信息熵等于最小值 0。图 1.1 直观地给出信息熵和 p_1 间的关系。相应地，随着 p_2 的变化，信息熵具有相同的变化。

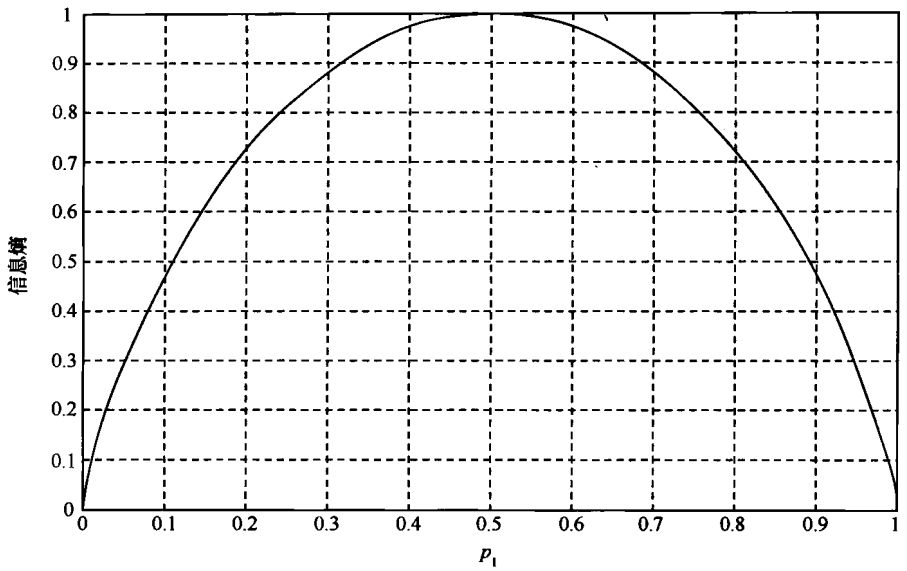


图 1.1 p_1 与信息熵的关系

1.1.4 信息熵在信息系统中的应用

对于给定的一个样例集合，信息熵可以反映集合中样例的类别不纯度。

假设一个集合仅包含两类样例，正例和负例，其中正例的比例为 p_+，负例的比例为 p_-，则该集合的信息熵为

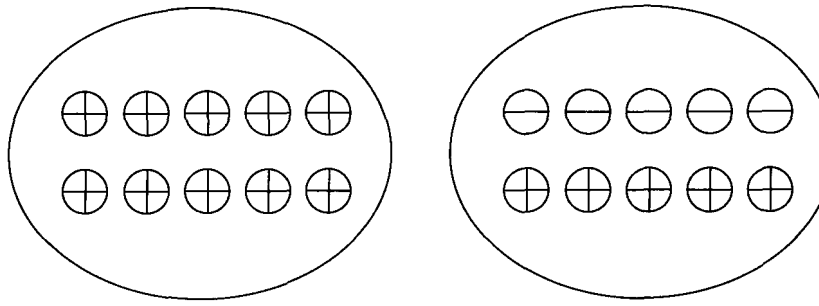
$$E(p)=-p_+\log_2 p_+-p_-\log_2 p_-$$

显然，当 $p_+=0$ 或 1 时，集合中的样例都属于同一类，集合的不纯度最小，集合的熵最小；当 p_+ 从 0 增大至 0.5 时，即随着正例比例不断从 0 增大到 0.5，集合的不纯度不断变大，而集合的熵逐渐增大；当正例和负例比例相等，即 $p_+=p_-=0.5$ 时，集合的不纯度最大、最混乱，此时集合的熵最大；当 p_+ 从 0.5 增大至 1 时，正例比例逐渐增加，集合的不纯度逐渐变小，即集合的熵逐渐变小；当 $p_+=1$ 时，集合中全部为正例，集合的熵达到最小值 0。因此，信息熵反映了集合中样例类别的混乱程度，即集合的不纯度大小。

图 1.2 所示的两个集合中，集合 A 的 10 个样例全部为正例，因此集合 A 中样例的类别分布为 $\left(\frac{10}{10},\frac{0}{10}\right)$，相应的其信息熵为 $E(A)=-1\cdot\log_2 1-0\cdot\log_2 0=0$；集合 B 中，5 个正例 5 个负例，其样例的类别分布为 $\left(\frac{5}{10},\frac{5}{10}\right)$，则集合 B 的信息熵为 $E(B)=-0.5\cdot\log_2 0.5-0.5\cdot\log_2 0.5=1$。显然集合 A 中样例全部为正例，

此为试读，需要完整PDF请访问：www.ertongbook.com

其信息熵最小，不纯度最小；而集合 B 中一半正例一半负例，其信息熵最大，样例类别分布最混乱，不纯度最大。



(a) 集合A　　(b) 集合B

图 1.2　集合的纯度和信息熵的关系

集合的信息熵还反映了集合能提供的信息量大小，即当随机地从集合中抽取一个样例时，该样例的类别具有不确定性。集合的信息熵越大，随机抽取的样例的类别越难以确定。若集合中的样例一半为正例一半为负例，则很难确定随机取到的样例类别；若集合中的信息熵较小，则抽取的样例的类别相对比较容易确定。若集合的信息熵为 0，则取出的样例的类别是确定的。比如，若集合中全部为正例，即 $p_+=1, p_-=0$，则取出的样例一定是正例。

相应地，对一个含有 N 个样例的集合 A，若集合中含有 L 类，各类样例数目分别为 $n_1, n_2, \cdots, n_L$，则集合的信息熵为

$$E(A) = -\sum_{i=1}^{L} \frac{n_i}{N} \log_2 \frac{n_i}{N}$$

显然，当集合中所有样例均为一类时，信息熵最小为 0，此时集合的不纯度最小；集合的熵变大，集合中样例不纯度逐渐增大，当集合中各类样例数目相等时，集合的信息熵达到最大，此时集合中样例类别分布最混乱，集合的不纯度最大。因此若随机地从集合中挑选一个样例，该样例的类别不确定性也最大。

1.2 模 糊 性

1.2.1 模糊现象

模糊性是人们在对客观世界认识过程中，无法给出清晰准确的界限而产生的不确定性[4]。生活中有许多模糊性概念，例如黎明、黄昏。这两个模糊概念是因为人们无法对昼夜转换给出一个清晰的界限而产生的。这些模糊性概念内涵很清楚，但外延模糊。我们每个人都明白黎明、黄昏的含义，但无法具体给出一个清晰

的时间段，说这个时间段就是黎明，那个时间段就是黄昏。在昼夜转化过程中，人们因为无法给出一个清晰的界限，所以在语言上产生了黎明和黄昏这两个模糊概念。模糊性是在人类对客观世界认知过程中产生的，因而这种模糊性更多地体现在人类的语言中。

在人类的认知世界中，有很多概念是模糊的，没有明确的两极边界，例如在日常生活中的大小、长短、轻重、高低等都是模糊概念。传统的集合理论很难对这类概念进行恰当地刻画。1965年，美国科学院院士Zadeh教授提出模糊集合论，为模糊概念的表示和模糊推理的形成奠定了数学基础。模糊集合论，使计算机可以跨越“黑白”两极边界，在“灰色”中间地带发挥作用[4]。

模糊集合论认为，论域上的对象从属于集合是逐渐过渡的，而不是突然变化的。它把元素属于集合的概念模糊化，认为论域上存在既非完全属于某集合，又非完全不属于某集合的元素；它又把属于概念量化，强调一个元素属于某一集合的程度，而不是集合中包含哪些元素。称元素属于某一集合的程度为隶属度。

1.2.2 模糊集

模糊集和隶属函数的定义如下[5]。

定义 1.2.1 设$\widetilde{A}$是论域U到$[0,1]$的一个映射，即

$$\widetilde{A}:U\to[0,1],\quad u\mapsto\widetilde{A}(u)$$

称$\widetilde{A}$为U上的模糊集，称$\widetilde{A}(u)$为模糊集$\widetilde{A}$的隶属函数。

从上述定义可以看出，模糊集$\widetilde{A}$完全由其隶属函数$\widetilde{A}(u)$刻画，把论域U中的每一个元素u都映射为$[0,1]$上的一个值$\widetilde{A}(u)$，$\widetilde{A}(u)$越大，表示元素u隶属于模糊集$\widetilde{A}$的程度越高。当$\widetilde{A}(u)$的值只取0或1时，模糊集$\widetilde{A}$便退化为一个普通集合（清晰集合）。

为了书写方便，在不引起混淆的情况下，后面将不加说明地省去模糊集$\widetilde{A}$上面的波浪符号，简写为A。

下面介绍模糊集合的表示方式。

(1) 序对表示法：$A=\{(u,A(u))\mid u\in U\}$。

(2) 若U为有限集$u_1,u_2\cdots,u_n$，则A可表示为

$$A=\frac{A(u_1)}{u_1}+\frac{A(u_2)}{u_2}+\cdots+\frac{A(u_n)}{u_n}$$

或者

$$A=\{A(u_1),A(u_2),\cdots,A(u_n)\}$$

称前者为Zadeh表示法，后者为向量表示法。

下面介绍模糊集合的基本运算。

定义 1.2.2 设 A,B 是论域 U 上的模糊集，分别称模糊集 $A\cup B,A\cap B$ 为 A 和 B 的并和交，而称模糊集 A^c 为 A 的补集。其中，$\forall u\in U$，则

$$(A\cup B)(u)=\max\{A(u),B(u)\}=A(u)\vee B(u)$$

$$(A\cap B)(u)=\min\{A(u),B(u)\}=A(u)\wedge B(u)$$

$$(A^c)(u)=1-A(u)$$

A 是 B 的子集定义为当且仅当对于所有的 $u\in U$，存在 $A(u)\leqslant B(u)$。

定义 1.2.3 模糊集合 A 的大小称为基数，记作

$$M(A)=\sum_{u\in U}A(u)$$

下面介绍常用的隶属函数。从理论上说，隶属函数的形式多种多样，但从实用角度来说，常用的隶属函数有三角隶属函数、梯形隶属函数和高斯隶属函数。

图 1.3 给出三角隶属函数的图形。

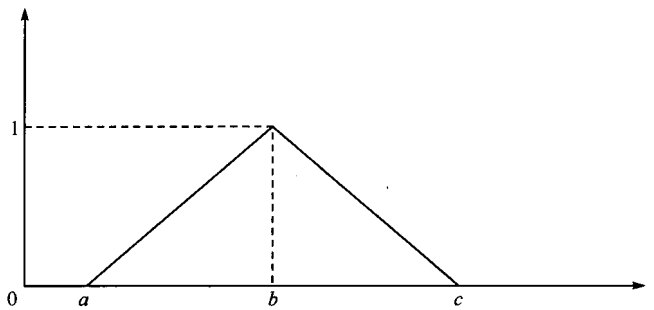


图 1.3 三角隶属函数

1.2.3 模糊性度量

模糊集合的隶属度刻画了一个元素属于集合的程度。一个模糊集合的模糊程度(不确定性)可以利用模糊熵[6~10]来度量，这个模糊熵类似于信息熵。

定义 1.2.4 模糊性度量：设 A 是论域 U 上隶属函数为 $\mu_A(u)$ 的模糊集合，若 $U=\{u_1,u_2,\cdots,u_m\}$ 是离散集合，并且 $\mu_i=\mu(u_i)$，则模糊集合 A 的模糊性度量为

$$E_V(A)=-\frac{1}{m}\sum_{i=1}^{m}[\mu_i\ln(\mu_i)+(1-\mu_i)\ln(1-\mu_i)]$$

$E_V(A)$ 度量了一个模糊集合的模糊度。当对每个 $u\in U$ 都有 $\mu_A(u)=0.5$ 时，$E_V(A)=1$ 最模糊；当对每个 $u\in U$ 都有 $\mu_A(u)=0$ 或 1 时，$E_V(A)=0$ 最清晰。

例 1.2.1 甲、乙、丙、丁四个人的身高分别为 1.8m、1.7m、1.75m、1.60m。模糊集合 T 表示“高个”这个模糊概念，假设四个人的身高隶属于 T 的隶属度分别为 1.0、0.5、0.7、0.1。试度量模糊集合 T 的模糊性。

解

$$
\begin{aligned}
E_V(A) &= -\frac{1}{4}\sum_{i=1}^{4}(\mu_i\ln(\mu_i)+(1-\mu_i)\ln(1-\mu_i)) \\
&= -\frac{1}{4}\times(0+0.5\ln 0.5+0.5\ln 0.5+0.7\ln 0.7+0.3\ln 0.3+0.1\ln 0.1+0.9\ln 0.9) \\
&= -\frac{1}{4}\times[0+(-1)+(-0.3602)+(-0.5211)+(-0.3322)+(-0.1368)] \\
&= \frac{1}{4}\times 2.3503 \\
&= 0.5876
\end{aligned}
$$

1.3 不可指定性

1.3.1 不可指定性现象

不可指定性是另一种认知不确定性。它与一对多的关系有关，例如在面对两个或更多选择形势下，无法给出明确的选择。这种不确定性在生活中也经常出现，例如，去商场买东西时，面对琳琅满目的商品，当有多件商品都差不多时，有时很难做出选择。

为了进一步说明不可指定性，再举一个例子。假设你是一个公司的招聘人员，有甲、乙、丙三个候选人，他们的考试成绩分别是86、84、85。现在让你进行选择，你可能感觉不好做出选择，因为他们的成绩很接近。另一种情况，假设他们的考试成绩分别是99、60、65。现在再进行选择，则可能很容易做出选择，因为甲的成绩比另外两人明显高出很多，后一种情况比前一种情况的不确定性要小。

不可指定性越大，人们就越难做出选择；不可指定性越小，人们就越容易做出选择。

1.3.2 可能性分布

可能性是人们对事物的可实现程度或达到某目标的难易程度的一种反映，其大小与人的主观判断有关。例如“这次实验可能成功”，“张三可能考上大学”等，都是对一种结果的不确定性的估计。针对这一问题1978年Zadeh创立可能性理论[4]。这里只简单介绍可能性分布。

定义 1.3.1 设映射 $\pi: X\to[0,1]$ 满足 $\bigvee\limits_{x\in X}\{\pi(x)\}=1$，则称 π 是 X 上的可能性分布函数[5]。

1.3.3 不可指定性度量

可能性分布[6,7]的不可指定性度量可以定义如下。

定义 1.3.2 不可指定性度量：设 $\pi=(\pi(x_1),\pi(x_2),\cdots,\pi(x_n))$ 为 $X=\{x_1,x_2,\cdots,x_n\}$ 上的正则可能性分布，则可能性分布 π 的不可指定性度量为

$$E_a(\pi)=\sum_{i=1}^{n}(\pi_i^*-\pi_{i+1}^*)\ln i$$

其中，$\pi^*=\{\pi_1^*,\pi_2^*,\cdots,\pi_n^*\}$ 是 π 的降序排列，即 $\pi_i^*\geqslant\pi_{i+1}^*$，且 $\pi_{n+1}^*=0$。

还以前面公司招聘为例来计算可能性分布的不可指定性度量。

(1) 某公司要从甲乙丙三个候选人中招聘一名职员，他们的考试成绩分别是86、84、85，试计算不可指定性度量。

(2) 假设他们的考试成绩分别是99、60、75，试计算不可指定性度量。

解 本例并没有给出正则可能性分布，因而需要先确定正则可能性分布，把成绩分别除以最高分即可得到正则可能性分布。第一种情况正则化后的分布为1，0.9767，0.9884，第二种情况正则化后的分布为1，0.6061，0.7576。然后进行排序，第一种情况排序后为1，0.9884，0.9767，第二种情况排序后为1，0.7576，0.6061。下面分别计算两种情况下的不可指定性度量。

(1)

$$\begin{aligned}E_a(Y)&=\sum_{i=1}^{4}(\pi_i^*-\pi_{i+1}^*)\ln i\\&=(1-0.9884)\times\ln1+(0.9884-0.9767)\times\ln2+(0.9767-0)\times\ln3\\&=0+0.0117\times\ln2+0.9767\times\ln3\\&=1.5597\end{aligned}$$

(2)

$$\begin{aligned}E_a(Y)&=\sum_{i=1}^{4}(\pi_i^*-\pi_{i+1}^*)\ln i\\&=(1-0.7576)\times\ln1+(0.7576-0.6061)\times\ln2+(0.6061-0)\times\ln3\\&=0+0.1515\times\ln2+0.6061\times\ln3\\&=1.1121\end{aligned}$$

对于第一种情况如果让你进行选择，你可能感觉不好做出选择，因为他们的成绩很接近；对于第二种情况，让你进行选择，你可能会很容易做出选择，因为甲的成绩比另外两人明显高出很多。通过计算可见，第二种情况下不可指定较小，人们更容易做出选择，这与人们通常的感受是一致的。

1.4 粗 糙 性

1.4.1 粗糙性现象

1904年Frege提出含糊(vague)一词，把含糊现象归结到边界线上。在论域

上存在一些个体不能在其某个子集上分类也不能在该子集的补集上分类。

1982年波兰数学家Pawlak针对Frege的边界线区域思想提出了对不确定知识进行表示的粗糙集(rough sets)理论[12]。下面通过一个例子来介绍粗糙性。

表1.1 粗糙性的例子

商店	质量	价格	盈利状况
1	好	低	盈利
2	好	高	亏损
3	好	高	盈利
4	一般	高	亏损
5	一般	高	亏损
6	一般	低	亏损

从表1.1可以得出下面的知识。

确定的规则:质量“好”且价格“低”,则盈利。例如商店1。

确定的规则:质量“一般”且价格“高”或“低”,则亏损。例如商店4,5,6。

可能的规则:质量“好”且价格“高”,可能盈利。例如商店3。

可能的规则:质量“好”且价格“高”,可能亏损。例如商店2。

在上面的例子中6个商店被分成了3个集合。

盈利:{1};

可能盈利可能亏损:{2,3};

亏损:{4,5,6}。

下近似集:{4,5,6}一定亏损;

上近似集:{2,3,4,5,6},上近似集的补集{1}一定盈利;

边界集:{2,3}上下近似集之差,可能亏损可能盈利。

在这个例子中可以用上下近似集来描述知识的不确定性,称为粗糙集。

用粗糙集来描述的不确定性称为粗糙性。粗糙性可以随着知识的增加而消失。例如在表1.1的基础上添加“经验”属性,如表1.2所示会发现亏损的上下近似集合相等,粗糙性消失。

表1.2 粗糙集的例子

商店	质量	价格	经验	盈利状况
1	好	低	丰富	盈利
2	好	高	缺乏	亏损
3	好	高	丰富	盈利
4	一般	高	缺乏	亏损
5	一般	高	丰富	亏损
6	一般	低	缺乏	亏损

此为试读,需要完整PDF请访问: www.ertongbook.com