

概率与模糊信息论 及其应用

刘立柱 编著

国防工业出版社

<http://www.ndip.cn>

内 容 简 介

本书总结了作者多年来在信息科技领域的科研成果、学术研究成果和教学经验,吸收了许多学者所发表的最新理论与技术成果,跟踪该领域发展前沿,具有较强的系统性、逻辑性和准确性。

全书分为4个部分,第1部分为概论;第2部分为概率信息论,包括第2章~第6章;第3部分为模糊信息论,包括第7章~第14章;第4部分,即第15章为概率信息和模糊信息的综合处理及其应用。书中较详细地讨论了概率熵和模糊熵理论,考虑到如何把模糊信息论应用于各个技术领域,详细阐述了模糊集的隶属度函数的模糊统计法、模糊德尔菲法、优化处理方法以及神经网络确定方法;从有利于培养学生分析问题解决问题的能力、工程应用能力以及工程技术人员参考出发,力求理论与技术的统一,包括了逼近码长下限的理论与技术、通信系统优化设计的理论与技术、信号的模糊分类与识别、信号模糊检测与模糊决策、图像的模糊处理方法和模糊信息系统等实用性内容。

该书既适合作为通信、计算机、信号与信息处理、遥感遥测、自动控制等专业的本科生和研究生教材,也可供科研人员和工程技术人员作为参考用书。

图书在版编目(CIP)数据

概率与模糊信息论及其应用 / 刘立柱编著. —北京:
国防工业出版社, 2004.7
ISBN 7-118-03483-5

I. 概... II. 刘... III. 熵(信息论) IV. 0236

中国版本图书馆 CIP 数据核字(2004)第 035945 号

国防工业出版社 出版发行

(北京市海淀区紫竹院南路 23 号)

(邮政编码 100044)

腾飞胶印厂印刷

新华书店经售

*

开本 787×1092 1/16 印张 18 $\frac{3}{4}$ 430 千字

2004 年 7 月第 1 版 2004 年 7 月北京第 1 次印刷

印数:1-3000 册 定价:25.00 元

(本书如有印装错误,我社负责调换)

前 言

21 世纪的社会是变革的社会,变革的总趋势是社会信息化。社会信息化是指以计算机与通信网络为主体的数字化、网络化、智能化和可视化的全过程。所谓数字化,是指信息的表示、传输、存储媒体的数字化,这有利于信息的处理、传输和应用。所谓网络化,是指数字化的信息通过高速网络进行流动,为社会各领域、各成员提供各种通信业务,用形象的说法就是形成了信息社会的神经系统。可视化,是指信息的直观表现形式,使人们更容易理解其意义。所谓智能化,是指对信息的处理、传输和利用的自动化,在计算机技术、通信技术、感测技术和控制技术领域都占有重要的位置。

社会信息化还应包括军事变革中的信息化。未来战争的战场形态将主要表现在以精确打击为主的非接触式、非线性作战样式,核心问题是争夺制信息权,主要是体系间的对抗。作战样式将是陆、海、空、天的联合作战。因此,对处于机械化半机械化的我军,如何尽快实现信息化是一个刻不容缓的战略任务。

从社会信息化和军队信息化建设需求出发,促使我对概率信息和模糊信息理论与技术开始研究,并且在为研究生和本科生讲授“信息论与信源编码”、“信息科学原理”、“模糊模式识别”、“模糊信息处理”等课程中,以及在指导研究生进行课题研究中,积累了一些研究成果和经验,这都促使我编著《概率与模糊信息论及其应用》这本书。

信息论的创始人香农(C. Shannon)研究了通信系统,在 1948 年发表了《通信的数学理论》,对信息给出了定量的规定,从而开辟了对信息的了解和研究,香农用概率方法给出了信源和信道的数学模型,并对通信问题进行了深入的研究,取得了辉煌的成就,这是世人所公认的。香农在 1948 年的论文中提出了“通信的基本问题是在某一点(终端——信宿)准确地或近似地再现从另一点(信息源)选择的消息。”这句话提出了两类问题:第一类问题是,从信源中如何选择消息?第二类问题是,如何把已选用的消息可靠地传出去?也就是说,既要解决传输的有效性,也要解决传输的可靠性问题。美国麦克埃里斯在《信息与编码理论》一书中,详细阐述了信道的容量代价函数理论,使通信系统优化设计目标由三维(有效性、可靠性和安全性)变为四维(有效性、可靠性、安全性和经济性)。

模糊性与随机性存在着本质的区别。所谓随机性是对事件的发生与否而言的,由于条件不充分,事件可能发生也可能不发生,即事件的发生存在一定的概率,但事物本身的含义是明确的,研究随机性的工具为统计数学。模糊性是指事件本身的含义是不明确的,但事件发生与否是明确的,研究模糊现象的数学工具是模糊数学。模糊信息论以信息技术领域为背景,所追求的目标是信息系统的智能化、武器系统的智能化。模糊信息论与香农信息论同属于语法信息的研究领域,但有着不同的研究对象、不同的研究工具、不同的研究内容、不同的应用环境和不同的研究目标。

《概率与模糊信息论及其应用》一书,从体系结构上看分为 4 个部分,第 1 部分为概

论;第2部分为概率信息论,包括第2章~第6章;第3部分为模糊信息论,包括第7章~第14章;第4部分,即第15章为概率信息和模糊信息的综合处理及其应用。第2部分和第3部分有平行的体系结构特点。编著本书时致力于系统性、逻辑性和先进性,强化基础、突出基本理论,较详细地讨论了概率熵和模糊熵理论,鉴于模糊集的隶属度函数在模糊理论中的核心地位,考虑到如何把模糊信息论应用于各个技术领域,给出了模糊集的隶属度函数的模糊统计法、模糊德尔非法、优化处理方法以及神经网络确定方法;从有利于培养学生分析问题解决问题的能力、工程应用能力以及工程技术人员参考出发,力求理论与技术的统一,包括了逼近码长下限的理论与技术、优化设计的理论指南与实践、雷达信号的模糊检测、图像处理的模糊方法和模糊逻辑系统等实用性内容。该书既适合作为本科生和研究生的教材,也可供科研人员和工程技术人员作为参考用书。

本书内容除总结了本人科研成果、学术研究成果和教学经验以外,也包含了我的研究生们的有关研究成果。需特别指出的是:本书中采纳了许多学者所发表的理论与技术成果,如参考文献中所列出的;宋国文老师曾审阅过第2部分的初稿,提出了许多很好的修改意见,在此一并对各位学者表示衷心的感谢。

由于作者水平所限,写作时间仓促,而且该领域的理论和技术仍在发展之中,书中如有不妥或错误之处,敬请读者批评指正。

刘立柱

于解放军信息工程大学

2003年12月28日

目 录

第 1 章 概论	1
1.1 信息的概念与分类	1
1.2 香农信息论与模糊信息论	5
1.3 智能系统与信息技术智能化	7
第 2 章 信源特性分析	10
2.1 离散信源的熵	10
2.2 条件熵与 Fano 不等式	14
2.3 微分熵和随机矢量的熵	17
2.4 互信息函数	20
2.5 数据处理定理	25
2.6 随机矢量之间的互信息	28
2.7 信源的冗余度	33
第 3 章 逼近码长下限的理论与技术	41
3.1 无失真信源编码分析	41
3.2 定长编码	43
3.3 变长编码	48
3.4 变长码的编码方法	54
3.5 游程长度(RL)编码	61
3.6 MH 码的编译码技术	66
3.7 二进制算术编码	71
3.8 文本压缩编码	74
第 4 章 信道的信息传输能力分析	79
4.1 信道容量	79
4.2 DMC 的容量代价函数	81
4.3 高斯信道的容量代价函数	88
4.4 信道编码定理	91
第 5 章 限失真条件下逼近码长下限理论与技术	99
5.1 引言	99
5.2 离散无记忆信源的速率失真函数	100
5.3 高斯信源的 $R(\delta)$ 函数	111
5.4 信源编码定理	115
第 6 章 优化设计的理论指南与实践	121

6.1	预测变换编码	121
6.2	正交变换编码	124
6.3	矢量量化与 $R(\delta)$	131
6.4	接入网优化设计的理论武器	135
6.5	扩展频谱通信的理论基石	141
6.6	通信系统优化设计理论与实例	143
第 7 章	模糊信息的数学模型	147
7.1	模糊信息与模糊集合	147
7.2	模糊集合的运算	150
7.3	模糊集合与普通集合之间的关系	156
7.4	凸模糊集和模糊数	158
第 8 章	隶属函数的优化求解原理与技术	162
8.1	确定隶属函数的一般方法	162
8.2	基于数据样点模糊划分的隶属函数求解方法	166
8.3	神经网络的基本知识	172
8.4	模糊集的隶属函数神经网络求解方法	174
第 9 章	模糊不确定度与模糊信息量	180
9.1	格和算子	180
9.2	测度	184
9.3	模糊集合的熵	190
9.4	模糊信息量	195
第 10 章	关联模糊信息处理准则与模糊聚类技术	197
10.1	关系的基本知识	197
10.2	模糊关系	198
10.3	模糊关系的合成	202
10.4	模糊关系的自反性、对称性和传递性	205
10.5	模糊聚类	208
第 11 章	模糊模式识别	217
11.1	基于模糊集隶属函数的模式识别	217
11.2	基于模糊集距离准则的模式识别	220
11.3	汉语语音模糊识别的处理方法	226
11.4	手写数字的识别	228
第 12 章	图像信息模糊处理技术	233
12.1	图像边缘增强的模糊方法	233
12.2	遥感图像数据融合	237
12.3	L 相二维模糊熵	239
第 13 章	模糊决策	242
13.1	概述	242
13.2	统计决策	244

13.3	基于隶属函数的模糊决策的基本理论·····	245
13.4	基于模糊数的模糊决策·····	248
13.5	基于模糊一致关系的模糊决策·····	253
第 14 章	随机模糊信息处理 ·····	257
14.1	模糊概率·····	257
14.2	模糊事件的均值与方差·····	260
14.3	模糊概率的应用·····	261
第 15 章	模糊系统原理与技术 ·····	263
15.1	概述·····	263
15.2	传真信号传输速率模糊识别系统·····	266
15.3	舰船雷达目标的模糊模式识别系统·····	274
15.4	雷达信号的模糊检测系统·····	280
15.5	模糊信息处理系统·····	283
参考文献	·····	290

第 1 章 概论

1.1 信息的概念与分类

信息论的创始人香农(Shannon)研究了通信系统,在 1948 年发表了《通信的数学理论》,对信息给出了定量的规定,从而开辟了对信息的了解和研究,香农用概率方法给出了信源和信道的数学模型,一般的通信系统模型如图 1.1 所示。

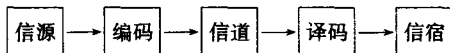


图 1.1 通信系统模型

所谓信源,就是任何一类在某个地方产生而在另一个地方感兴趣消息的设备。例如一本书、一张布告,以及舞蹈、音乐等都导致了多种多样的信息表现形式,它们本身就是信息源。

1. 信源的描述与分类

为了深入研究对信源符号的处理、编码等问题,我们首先讨论各种信源的描述,即信源的数学模型,进一步研究信源的数字特征和物理特性。比如掷一颗均匀骰子,每掷一次出现哪个点数是不可预测的,亦就是说是随机的。因此,该试验可用随机变量 X 表示试验结果,其值域 $R = \{1, 2, 3, 4, 5, 6\}$,每个数值出现的概率为 $p_i = 1/6 (i = 1, 2, 3, 4, 5, 6)$ 。该试验信源的数学模型可由概率空间 $[R, P]$ 表示

$$[R, P] = \begin{pmatrix} 1, & 2, & 3, & 4, & 5, & 6 \\ p_1 & p_2, & p_3, & p_4, & p_5, & p_6 \end{pmatrix}$$

因此,我们以后所研究的随机变量、随机序列,以及随机过程都可作为信源的统计模型。这一点请注意。

(1) 离散信源 就是随机变量的值域 R 为一离散集合。记 X 的值域 $R = \{x_1, x_2, \dots, x_n\}$ 。要充分描述 X ,只要规定了 R 集合中各个元素的概率测度 $p(x_i)$,也就规定了所有 R 集合的波雷尔集合的概率测度。一般的 X 的值域 R 和 R 集合的概率分布被称之为概率空间,并简记 $[R, P]$ 。

$$[R, P] = \left[\begin{matrix} x_1, & x_2, & \dots, & x_n \\ p(x_1), & p(x_2), & \dots, & p(x_n) \end{matrix} \right], \sum p(x_i) = 1$$

(2) 连续信源 即出现的消息数是无限的或不可数的。可用连续性随机变量来描述,其数学模型为

$$[R, p(x)] = \left[\begin{matrix} R \\ p(x) \end{matrix} \right] \int p(x) d(x) = 1$$

式中 $R=(a,b)$ 或 $R=(-\infty, \infty)$, $p(x)$ 为连续随机变量 X 的概率密度函数。

应该指出的是:实际上,信源发出的不一定都是一个一个的信源符号,而是一系列的信源符号,更为一般的是时间的函数 $X(t)$,通过抽样可得到时间上离散的诸个随机变量,可用随机序列表示,即

$$X_1, X_2, \dots, X_N \text{ 或 } X_1, X_2, \dots, X_h, \dots$$

若 $X=(X_1, X_2, \dots, X_N)$, 当 X_h 的值域 $R=\{x_1, x_2, \dots, x_n\} (h=1, 2, \dots, N)$ 。那么

$$x=(x_1, x_2, \dots, x_N) \in R^N$$

$p(x)=p(x_1, x_2, \dots, x_N)=p(X_1=x_1, X_2=x_2, \dots, X_N=x_N)$ 。可见 X 的取值有 n^N 种,若对每一种 X 的取值 x 规定了概率测度,也就充分描述了随机变量 X 。

(3) 无记忆信源与有记忆信源 当 X_h 相互独立时,且有同样的概率密度函数,则有 $p(x)=\prod_{h=1}^N p(x_h)$, 与此相对应的信源被称为无记忆信源;反之,就是有关联的,与此相对应的信源被称为有记忆信源。该信源的描述相对比较困难,因此在实际问题中往往限制记忆的长度,即 X_h 的概率密度仅仅与它前面的一个(X_{h-1})或 M 个($X_{h-1}, X_{h-2}, \dots, X_{h-M}$)有关,当只与 X_{h-1} 有关时,被称之为马尔柯夫链性质,这样的信源就是有记忆的信源。

(4) 平稳性信源与非平稳信源 就是指信源发出的消息序列 $X_1, X_2, \dots, X_h, \dots$ 具有下述特性:

① 信源在 $t=h$ 时刻发出什么符号与信源在 $t=h$ 时刻随机变量 X_h 取 x_h 符号的概率分布 $p(x)$ 有关。一般情况下,在 $t=i$ 时的 $p(x_i) \neq p(x_h)$ 。

② 信源在 $t=h$ 时刻发出什么符号与信源在 $t=h$ 时刻以前发出的符号有关,即与 $p(x_h/x_{h-1}, x_{h-2}, \dots)$ 有关。一般情况下, $p(x_h/x_{h-1}, x_{h-2}, \dots)$ 为时间 t 的函数,即

$$p(x_h/x_{h-1}, x_{h-2}, \dots) \neq p(x_i/x_{i-1}, x_{i-2}, \dots)$$

这样的信源被称为非平稳信源。而把满足

$$p(x_i) = p(x_h) \text{ 和 } p(x_h/x_{h-1}, x_{h-2}, \dots) = p(x_i/x_{i-1}, x_{i-2}, \dots)$$

的信源称之为平稳信源。它的统计特性与时间的起点无关。平稳信源发出的符号序列中,不同符号组的联合概率应满足

$$\begin{aligned} p(x_h, x_{h+1}) &= p(x_h)p(x_{h+1} | x_h) \\ p(x_h, x_{h+1}, x_{h+2}) &= p(x_h)p(x_{h+1} | x_h)p(x_{h+2} | x_h, x_{h+1}) \\ &\vdots \\ p(x_h, x_{h+1}, \dots, x_{h+N}) &= p(x_h)p(x_{h+1} | x_h) \cdots p(x_{h+N} | x_h, x_{h+1}, \dots, x_{h+N-1}) \\ &\vdots \end{aligned}$$

且有

$$\begin{aligned} p(x_h + N | x_h + N - 1, \dots, x_h + 1) &= \\ p(x_i + N | x_i + N - 1, \dots, x_i + 1) &= \\ p(x_N | x_{N-1}, \dots, x_1, x_0) & \end{aligned}$$

(5) 随机过程信源 即信源输出可用随机过程来描述的信源。此时,随机变量 $X(t)$

是时间 t 的函数, $t \in T$, $T = R$ 或 $T = (a, b) \in R$, R 为实数集。比较简单的过程有限时、限频过程。所谓限时过程, 就是随机过程 $X(t)$, $t \in (a, b)$; 所谓限频过程, 是指 $X(t)$ 具有频谱函数 $H(f)$, 亦就是满足

$$H(f) = 0, |f| > F$$

这两种过程都可以展开成傅里叶级数, 当然展开式系数之间一般还是线性相关的。对于限时随机过程也可以进行 $K-L$ 变换, 此时的展开式系数之间是相互独立或线性无关。

然而, 对于满足 $p(x_i) = p(x_h)$ 的信源则称为平稳性信源。否则, 就是非平稳性信源。

(6) 信道 指用于传输信息的任何一类物理媒介, 或者说是载荷着信息的信号所通过的通道(途径)。它的作用是传输信息和存储信息。它是以信号特别是电信号的形式载荷信息的, 例如一条电缆、一束光等。

2. 信道的描述与分类

要规定一个信道, 一是必须规定信道所容许的输入信号; 二是要确定输入、输出之间的统计依赖关系。这种依赖关系一般不是确定的函数关系, 而是统计依赖关系。可用条件概率来描述。信道的输入信号就是信源的输出符号, 前面已经进行了讨论, 而信道也可分为离散与连续、平稳与非平稳、有记忆与无记忆的。

应该指出: 这里的信道是广义的, 凡是讨论信息流通过路径中某一段的信号的统计依赖关系, 都可把这一段路径看作为信道来讨论。比如说, 纠错编码器的输入和输出之间统计依赖关系, 纠错编码器被看作为编码信道。因此, 信道按照输入、输出信号的取值, 可分为离散信道、连续信道、数字信道、半数字信道以及半连续信道。

信道的输入、输出之间的统计依赖关系可用条件转移概率进行描述, 其输入记为 X , $x \in A$, 那么, $X = (X_1, X_2, \dots, X_n)$, $x \in A_n$; 输出为 Y , $y \in B$, 而 $Y = (Y_1, Y_2, \dots, Y_n)$, $y \in B_n$ 。 x 与 y 之间的转移概率可由矩阵表示

$$[p(\bar{y} | \bar{x})] = Q$$

若 $p(\bar{y} | \bar{x}) = \prod_i p(y_i | x_i)$, 则该信道就是无记忆的; 否则就是有记忆的。

若 $p(\bar{y} | \bar{x}) = \{1, 0\}$, 则该信道被称之为无干扰或无噪声信道; 反之, 就是有噪声的。

对一个信道而言, 当输入、输出都不是单个的, 被称之为多用户信道(如多路复用信道)、多输出信道(如广播信道)等。多用户信道的最大传信率不是一个数值, 而是一个数值集合, 即容量区域。

信道的描述形式, 还有线图表示法和框图表示法。线图表示形式如图 1.2 所示。

框图表示法, 如离散无记忆信道 DMC, 其输入为 $\{0, 1, 2, \dots, r-1\}$, 其输出为 $\{0, 1, 2, \dots, s-1\}$, 该信道的框图表示形式如图 1.3 所示。

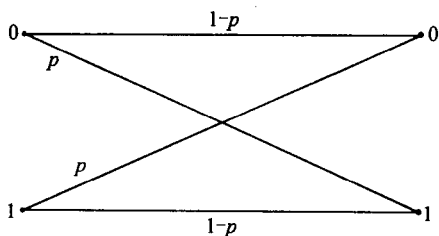


图 1.2 信道的线图表示

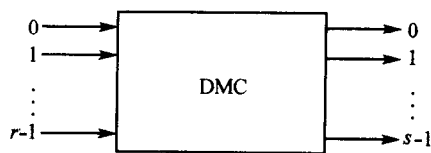


图 1.3 信道的框图表示

编码,实际上就是用符号来表示信息,或者说是用适合于信道传输的符号来表示信源输出的符号。而译码是编码的逆变换。

香农在 1948 年的论文中提出了“通信的基本问题是在某一点(终端—信宿)准确地或近似地再现从另一点(信息源)选择的消息。”这句话提出了两类问题:第一类问题是,从信源中如何选择消息?第二类问题是,如何把已选用的消息可靠地传出去?也就是说,既要解决传输的有效性,也要解决传输的可靠性问题。由此可见,对信源而言核心问题是它包含的信息到底有多少?这就是信源熵函数问题。而信道的核心问题是它最多能传送多少信息?这就是信道的容量问题。在一定的限制条件下,就是容量—代价函数问题。如何编码才能使信息源的信息被充分表示、信道容量才能被充分利用,这些编、译码方法是否存在?这是编码定理所要解答的问题。在牺牲一定的保真度的条件下,具有更大压缩比的编、译码方法是否存在?或者说压缩与失真如何折中?这就是速率失真函数问题。总之,香农信息论是研究通信系统的可靠性、有效性和低成本问题,是寻求系统优化的基本理论。

在 20 世纪后期,以信息作为主要研究对象、以信息的运动规律为主要研究内容、以现代科学方法论为主要研究方法、以扩展人的信息功能(尤其是其中的智力功能)为主要研究目标的一门科学诞生了,被称为“信息科学”。它所研究的运动规律包括信息产生规律、信息提取规律、信息再生规律、信息施效规律。利用信息科学方法论研究在机器上复现主体的提取信息的功能、传递信息的功能、处理信息的功能、再生信息的功能、施用信息的功能,尤其是这些功能的综合亦即智力功能,就是信息科学的研究目标。

目前信息科学的研究重点可概括为:揭示利用信息来描述系统和优化系统的方法和原理;寻求通过加工信息来生成智能的机制和途径。

能够扩展人的信息器官功能的技术群,称为信息技术。它包括:感测技术——感觉器官功能延伸;通信技术——传导神经功能延伸;计算机技术——思维器官功能延伸;控制技术——效应器官功能延伸。应该指出:通信技术和计算机技术是信息技术的核心,感测技术和控制技术的核心与外部世界的接口,信息技术四基元是一个完整的体系。因此可以说信息技术是指遥感遥测、通信、计算机、控制等技术的整体。

何谓信息?有人说,“信息就是消息”、“通信的内容就是信息”。这些说法不够严谨,信息与消息是有区别的,信息是消息或消息传输的概括,消息是信息的具体化。所谓信息必须具有“能消除某些知识的不肯定性”的秉性。就是说能改变人们的知识状态,使从无知变为有知、从不肯定变为肯定。广义信息是指人类感官所能感知(直接的或间接的)的一切有意义的东西。比如,电话、电视、雷达、声呐可以给我们传来信息。意大利学者 G. Longo 在 1975 年出版的《信息论:新的趋势与未决问题》一书序言中,称“信息是反映事物的形式、关系和差别的东西,它包含在事物的差异之中,而不在事物本身”。我国已故电子学家冯秉铨先生也赞同“信息就是差异”的理解。钟义信教授在他的《信息科学原理》一书中对信息进行了精辟的阐述,下面就给出他的论述。

关于信息的定义。

① 本体论层次的信息:就是事物运动的状态和状态改变的方式。

② 认识论层次的信息:就是认识主体所感知或所表述的事物运动的状态和状态改变的方式。

③ 语法信息:就是认识主体所感知或所表述的事物运动的状态和方式的形式化关系。

④ 语义信息:就是认识主体所感知或所表述的事物运动的状态和方式的逻辑含义。

⑤ 语用信息:就是认识主体所感知或所表述的事物运动的状态和方式相对于某种目的的效用。

关于语法、语义和语用信息的含义,可通过一个例子来进一步说明。例如爱因斯坦的著名的物理学公式, $E = mc^2$ 给出了能量与质量之间转换关系的信息。从语法信息的角度看,这个公式规定了 E 、 m 、 c 和数字 2 的一种排列方式,在没看到该式之前我们不知道哪几个字母、哪个数字会出现,也不知道它们以何种方式排列,具有一定的不定性。而在观察到该公式时,我们得知了它们的“存在状态和方式”,消除了先前的不定性而获得了语法信息。然而若不知道这些字母、数字和它们的排列方式有何含义,可以说我们并不知该式的语义何在,当得知 E 代表能量、 m 代表质量、 c 代表光速、2 代表平方关系时,才知道了这个公式的含义,消除了先前的语义不定性而获得了语义信息。对于观察者是否有价值?那要看是做什么工作的,比如对搞核能开发者而言,根据这一公式就可以通过改变原子核的质量状态而获得巨大的原子核能。

另一个例子是:香农著名的限带限功率高斯信道容量公式 $C = W \log(1 + P/NW)$,当我们知道了 W 为信道带宽、 P 为信号功率、 N 为噪声功率谱密度以后,也就是知道了该式的逻辑含义,才由此得知当信道容量一定时,带宽与信噪比可以互换,由此发展了扩频通信技术。这个公式对通信科技工作者而言显然是非常重要的。扩频通信技术是通信领域中一个发展迅速的技术分支,扩频通信是一种信号隐藏传输技术,即是把信号淹没在噪声之中,在军事通信中占有重要的位置,这个公式没有给出具体的实施方案,而是指明了一个方向,这就是科学理论的意义所在。

应该指出:语法信息仅涉及“事物运动的状态和状态改变的方式”,是最基本的层次。大家知道,在语言学里只考虑“词与词的结合方式”的研究称为语法学。在此我们把只考虑“状态和状态改变的方式”这一层次的信息称之为语法信息。同理,借用语言学中的语义学这个术语把只考虑“状态和状态改变的方式的含义”这一层次的信息称之为语义信息。借用语言学中的语用学术语,而把只考虑“状态和状态改变的方式的含义的效用”这一层次的信息称之为语用信息。

目前大家所研究的是一些基本信息,例如语用信息、语义信息、概率信息、偶发信息、确定型信息和模糊信息。

在语法信息中,根据运动状态和方式的不同可细分为概率信息、偶发信息、确定型信息和模糊信息。概率信息的运动状态是完全按照概率规则和统计规律出现相联系的。偶发信息是指:各状态的出现是随机的而不是确定性的,但是由于这类试验只能进行若干次而不可能大量重复,不能用概率统计规则来描述,因此这类实验所提供的信息称之为偶发信息。确定型的运动方式是指其各状态的出现规则能用经典数学公式来描述,而未知因素表现在初始条件和环境影响方面,与这类运动方式相对应的信息称之为确定型信息。

1.2 香农信息论与模糊信息论

在信息科学发展中,模糊信息理论是目前值得注意的一个方向,它与香农信息论有

着本质的区别。在此首先介绍几个基本概念。

(1) 模糊性 指客观事物的差异在中介过渡时呈现的“亦此亦彼”性。如美与丑、清洁与污染、图像的清晰与模糊、音质好与坏等。这些对立的观念都没有分明的界限,即无明确的外延。称这些无明确的外延的念为模糊概念。因此,模糊性也就是无法确定其界限的性质。

模糊概念可以用集合来描述,即集合可以表现概念,一个概念有其内涵与外延。内涵指的是符合此概念的对象所具有的共同属性,外延指的是符合此概念的那些对象的全体。模糊理论不对事物做简单的肯定与否定,而是用隶属度来反映某一事物属于某范畴的程度,用这种方法来表示客观存在的模糊性。

例 1.1 关于人,“年轻”这个概念。究竟多大年龄以下为“年轻”,这个标准是以每个人的主观感觉为依据的,很难划定。我们不妨把年轻到什么程度的问题,用 0~1 之间的数来表示,如 15 岁~40 岁分别表示成:15→1.0,20→0.9,25→0.8,30→0.6,35→0.4,40→0.2。这就是一个模糊集合。由于这些程度是构成该集合的元素,因此使它们对应起来的函数就称之为隶属度函数。

例 1.2 某产品的质量共分为三个等级,以 a 表示一等品,出现的概率 $p(a)=0.5$;以 b 表示二等品,出现的概率 $p(b)=0.4$;以 c 表示三等品,出现的概率 $p(c)=0.1$ 。在这里出现一等品显然是一个普通的随机事件,而一等品的概念是明确的,有具体的检测标准。但是若问“质量好的产品占多大比例”?对此而言,“质量好的产品”本身就是一个模糊概念,这时我们只好用一等品属于“质量好的产品”这一子集的隶属度为 1,二等品属于“质量好的产品”这一子集的隶属度为 0.5,三等品属于“质量好的产品”这一子集的隶属度为 0 来对其进行描述。

可见模糊性的东西本身就是没有明确的界限。

(2) 模糊信息 我们首先承认由于存在模糊性,就必然存在着某种不确定性。比如在传真通信中,一张本来黑白分明的传真图像,由于某种原因变得模糊不清了,对于那些半白半黑的灰度色调究竟应算作白还是黑呢?这就产生了不定性,而为了消除这种不定性就需要信息。因此我们就把与事物的模糊性相联系的信息称之为模糊信息。也可以说,模糊信息是以模糊状态显现出的一种表现形式。

(3) 模糊信息论 指研究模糊信息的本质、模糊信息的度量、模糊信息的处理、模糊模式识别、模糊检测、模糊决策、模糊信息优化处理等领域的理论。研究模糊现象的数学工具是模糊数学;研究的目的是构筑智能系统;应用领域为信息技术的四大领域,即通信、计算机、控制和感测技术。

应该指出:模糊性与随机性是存在本质区别的,所谓随机性是对事件的发生与否而言的,由于条件不充分,事件可能发生也可能不发生,即事件的发生存在一定的概率,但事物本身的含义是明确的。例如,抛掷硬币,国徽朝上与否无法确定,是随机的,但国徽含义是明确的。我们可以通过多次抛掷得出国徽朝上的概率。研究随机性的工具为统计数学。香农信息论以通信问题为背景,所追求的目标是通信系统的有效性、可靠性和经济性。模糊性是指事件本身的含义就是不明确的,但事件发生与否是明确的。例如“老张的病不轻”,老张有病是确定的,但老张病重到何种程度却是不明确的。研究模糊现象的数学工具是模糊数学。模糊信息论以信息技术领域为背景,所追求的目标是信息系统的智能化

或智能化水平的提高。可见模糊信息论与香农信息论同属于语法信息的研究领域,但有着不同的研究对象、不同的研究工具、不同的研究内容、不同的应用环境和不同的研究目标。

1.3 智能系统与信息技术智能化

模糊信息理论是一种基于模糊集理论的信息科学,是一个迅速发展的信息科学的分支。

模糊信息理论应成为模糊数学理论与工程应用的桥梁,成为指导工程实践的又一个有普遍意义的强有力的工具。在通信、计算机、声呐、雷达、导航、制导、空间测控、工业控制等各种电子系统中,模糊信息的提取、处理和利用占有极为重要的地位。模糊信息理论研究领域:

- (1) 研究模糊信息的本质;
- (2) 探讨人脑处理模糊信息的机制;
- (3) 研究高效能的模糊信息提取、处理和利用技术、构筑智能化的信息系统。

模糊信息理论在信息技术中有广泛的应用,从一些重要的研究中就可清晰可见。例如,LIFE——国际模糊工程研究所(1989年,在日本通产省倡导下,由一些企业和大学组成),下属的三个研究室,第一研究室:研究模糊控制,第二研究室:研究模糊智能信息处理,第三研究室:研究模糊计算机。其中第二研究室有三个研究课题:一是决策支持系统。从语义学的高度研究模糊,主体是图像理解,追求的目标是从自然遗传因子抽取人的图像,并对其语言予以说明。二是模糊专家系统,也有人称为模糊知识库系统。从上述研究中我们可以看到,模糊信息处理研究具有很大的吸引力。

我国在该领域有不少研究成果。如谢维信教授,国家自然科学基金重大项目——模糊信息处理、思维决策与机器智能;邹开其教授,国家自然科学基金重大项目——桥梁损伤评估及对策专家系统等。

新的研究方向,是把模糊信处理和神经网络理论相结合。由于多媒体信息处理的发展,人机对话使机器具有形象思维能力需求的增长,模糊理论恰恰可直接处理像概念、语言值一类信息,可实现将人的定性思维和判断方法定量化以适应计算机处理的过程。而模拟人脑智能特点和结构的人工神经网络能处理信号并具有自适应学习、联想记忆等功能。二者的结合必将发挥出更好的性能,对电子产业和信息科学的发展具有不可低估的作用。

模糊信息理论在信息技术中具体应用,比如在计算机领域,模糊信息处理是模糊计算机的工作基础;在控制领域已发展了模糊控制和智能控制;在通信领域中,智能化信息网(能够完成全部信息功能的网络,“全部信息功能”——即获取信息、处理信息、再生信息、传递信息、施用信息)于1999年由钟义信教授在计算机及网络的应用与发展学术会议上提出,在这篇论文中,提出了“信息的基础理论、网络的体系结构……网络的智能化服务六大研究领域,并指出:“信息网络的根本任务和生命力,在于能够为各种各样的用户提供越来越多和越来越好的智能化网络服务,实现工业生产、农业生产、国防安全、交通运输、商业贸易、科学研究、文化教育、医疗卫生等各领域的智能化和现代化。因此,研究信息网络

的智能服务是一个永恒的主题,而且是越来越紧迫的主题。”显然也少不了模糊技术在遥控、遥感、航天等领域的应用。可以用一句不很确切的话来描述是,“信息技术中无处不模糊”。

从目前看,社会信息化已是大势所趋,社会各领域都需要信息,发展趋势可概括为:数字化、网络化、智能化、可视化。

智能化,是指信息科学与技术的发展,使人们有能力在机器上复现人(主体)的那些信息过程的机制。它一方面导致大量高级智能信息系统的问世,另一方面也孕育着智能科学与技术的诞生,有的专家断言在 21 世纪的技术变革中,信息技术发展方向将是智能化。

为了阐述智能与知识的关系,我们首先说明何谓智能?我国出版的《辞海》中是这样描述的:“智能是人认识客观事物并运用知识解决实际问题的能力,集中表现在反映客观事物深刻、正确、完全程度上和应用知识解决实际问题的速度和质量上,往往通过观察、记忆、想像、思考、判断等表现出来。”而福格尔(Fogel)等人提出:“智能是一种有目的以某种好的方法使用某个有用信息的能力。”我国钟义信教授把智能定义为:智能,即是获取信息、处理信息、利用信息的能力。刘增良博士则把智能表述为:智能,就是能获取存储知识并运用知识解决实际问题的能力。至此,不难看出:知识是智能行为的基础,智能是一种能力,是一种学习能力、思维能力、分析问题和解决问题的能力、认识世界和改造世界的能力。应该指出:智能是一种“动态”行为,特别是一种“思维”行为,是一种知识和经验的综合运用过程。可以说,“思维”是智能行为的核心。直感和灵感也是一种思维活动,它与逻辑思维、形象思维不同的是速度上的差异,直感比灵感来得迅速。至此我们可以说:知识是智能的基础,智能是知识的提升物,是对知识加工、变换和利用的产物。

通过上述分析,我们认为一个智能系统应具有:合理的物理结构、完善的知识系统、健全的思维机制。一个智能系统的功能主要是一种“信息加工”功能,其智力即“信息加工”的速度和能力。

我们也可以认为一个智能系统就是可实现一种信息“映射”。可将智能系统分为两类。

第一类智能系统,是系统的输出和输入信息之间不保持固定不变的简单对应,而是在一定的“知识平面”内“线性”可变,这类系统在运用固有特性和本能进行信息处理时,有一定的灵活性和适应性。这是一类“适应性”系统,可称为适应性智能系统。也被称为低级智能系统。

第二类智能系统,具备了以主观能动性为特征的有意识的自学习和创造功能,具有灵活运用其本身固有的特性和本能的“能力”,也就是说本身也不是相对稳定的了。这类系统可称为意识型能动性智能系统。也被称为高级智能系统。

应该指出的是:智能系统首先应具有“适应能力”,即智能体对外部信息不但具有感知能力,还具备经过“思维”,能灵活做出“适应性”反映的能力。第二是应具有“学习能力”,否则它就无法“获取”作为信息处理基础的“知识”、也无法改进其“思维能力”。正因如此,一个智能系统,当且仅当具有“学习功能”和“思维功能”时才可称为智能系统。

美国扎德(L. A. Zadeh)教授在接受 1989 年本田奖仪式上的讲话时说:“……我认为模糊理论今后在两个领域取得较大进展。一是熟练技术者替代系统,这种系统将人无意识进行的操作由机器替代,如日本仙台市营地铁的自动驾驶系统。另一个是替代专家的

专家系统,像山一证券公司的股票交易系统及医疗诊断系统。为使专家头脑中所进行的思考与决策实现自动化,模糊理论将起重要作用。当然,模糊理论并不能解决所有各种问题,但是只要不回避现实中的不确定事物并加以认真对待,就有可能大大提高在不确定(模糊)环境中进行智慧思考与决策的人及机器的能力。”

日本明治大学信息科学中心所长、工学博士,从事模糊理论、多值逻辑和安全技术研究的向殿政南教授在1990年6月,趣谈“模糊”时指出:“要想发展智能工业,必须能输入模糊信息,特别是人类知识大多是使用语言表达的,不能不使用模糊理论。”

综上所述,自从1965年扎德教授提出“模糊集”理论之后,模糊数学得到了迅速的发展,模糊信息理论、模糊控制理论、模糊计算机理论等也已有多年的发展历程,同时信息科学的普及与发展,人工智能理论的发展,神经网络、专家系统的发展,其目标都是为了发展智能工业。21世纪信息技术的发展,必将对智能科学与技术提出更高的需求,因此,我们对“智能科学与技术”应给予足够的重视。关于智能科学与技术,从目前来看,应包括模糊信息处理理论、人工智能理论、模式识别、模糊逻辑与智能控制、神经网络、专家系统等研究领域。我认为,在今后的信息技术领域,对智能信息技术,其信息的加工和处理将由集中型向分布型转变,也是说智能技术将渗透到传输、控制、计算、感测等领域。

第2章 信源特性分析

信源,即信息的产生源。信源和信道特性是我们所关注的特性之一,因为它与信源编码关系密切且极其重要。为此,我们在这一章首先对各种信源的信息特性进行讨论,主要包括信源熵,即 $H(X)$ 、 $H(\mathbf{X})$ 、 $h(X)$ 、 $h(\mathbf{X})$ 等函数,讨论它们的基本概念、主要特性以及它们的物理意义。

2.1 离散信源的熵

设信源 X , X 的值域 $R = \{x_1, x_2, \dots\}$ 是有限的或可数的,令 $p_i = p\{X = x_i\}$ 。这就是说信源输出的信号是随机的,我们在未观测到信号之前,对信源发出什么信号是不能肯定的,否则通信就没有什么意义了。而通信的目的就是要想办法使接收者把它肯定下来,那么信号中到底具有多大的不肯定度? 即判断某一随机事件是否发生的难易程度如何? 接收者要收到多少信息才能把信号肯定下来? 下面就来讨论这些问题。

若信源概率空间为

$$X = \left\{ \begin{array}{cccc} x_1 & x_2 & \cdots & x_n \\ p_1 & p_2 & \cdots & p_n \end{array} \right\}$$

其不肯定度为

$$H(X) = H(p_1, p_2, \dots, p_n) = - \sum_{i=1}^n p_i \log p_i \quad (2.1)$$

这个公式是信息论中一个最基本的公式,给出这个公式的目的是使 $H(X)$ 有定量值。要定量一定要有单位。我们就把该式等于 1 的情况看作为 $H(X)$ 的单位。为此我们先设

$$X = \left\{ \begin{array}{cc} x_1 & x_2 \\ 0.5 & 0.5 \end{array} \right\}$$

那么

$$H(X) = - \sum_{i=1}^n p_i \log p_i = \frac{1}{2} \log 2 + \frac{1}{2} \log 2 \quad (2.2)$$

显然当对数底为 2 时,式(2.2)恰等于 1。若以 e 为底时, $H(X) = 0.693147$ 。可见式(2.2)之所以等于 1, 一是因为两个事件的概率相等,二是必以 2 作为对数的底。我们把这样得到的 1 叫做 1 bit(比特)。不严密地说,式(2.2)可解释为“两个事件的概率相等时,其不肯定度为 1 bit”。

两个等可能事件的概率空间是最简单而又经常遇到,将它的不肯定度作为单位是合适的。这是因为一个 N 进制的波形总可以用若干个二进制波形来表示。同时以 2 为底时运算也简单方便,而在数字通信中又常是以二进制传输。对于多进制而言,比如说