第3章 智能计算与识别理论基础

本章给出多源信息融合赖以发展的智能计算与识别理论基础，大部分内容比较新。

3.1 概述

3.1.1 模式识别的一般概念

模式识别是人类自然智能的一种基本形式。所谓“模式”就是指人类按照时间和空间中可观察的自然属性和认识属性对客观事物的划分。所谓“模式识别”就是依据某些观测获得的属性把一类事物和其他类型的事物区分的过程。例如，人们依据视觉、听觉、触觉等感官所接收的信息，能够正确认识外界事物，并能把一类事物与其他事物正确区分。

然而，我们现在研究的“模式识别”主要属于人工智能的范畴。随着20世纪40年代计算机的出现以及50年代人工智能的兴起，一种以计算机为工具进行“模式识别”的理论和方法便发展起来了，并迅速成为一门新的学科。

目前，模式识别的理论研究主要集中在两方面：一是关于生物体(包括人)感知客观事物机理的研究，这是生物学家、生理学家、心理学家、神经生理学家等的研究内容，属于认知科学的范畴；二是针对给定的任务，关于计算机实现模式识别的有关理论和方法研究，这是数学家、信息学家和计算机科学家的研究内容，属于应用信息科学的范畴。

用计算机实现的模式识别系统，一般由三个相互关联而又有明显区别的过程组成，即数据生成、模式分析和模式分类。所谓数据生成是将原始信息转换为向量，成为计算机易于处理的形式；所谓模式分析是对数据进行加工，包括特征选择、特征提取、数据维数压缩和决定可能存在的类别等；模式分类则是利用模式分析所获得的信息，对计算机进行训练，从而根据判别准则，以实现模式的分类。

现在通用的模式识别方法有两种，即统计模式识别方法和结构(句法)模式识别方法。所谓统计模式识别方法，就是对模式的统计分类方法，即结合统计概率论的Bayes决策系统进行模式识别的技术，

又称决策理论识别方法。利用模式与子模式分层结构的树状信息所完成的模式识别工作，就是结构模式识别或句法模式识别。

模式识别系统已经成功地应用于字符识别、语音识别、指纹识别、人脸识别、医学图像分析、工业产品检测等许多方面。然而，模式识别方法的研究目前仍然是针对具体应用对象的，至今还难以发展成为一种统一的模式识别理论与方法。

3.1.2 智能学习与统计模式识别

统计模式识别的基本原理，是根据先验知识，定义**模式空间**为

$$\Omega = \{\omega_1, \omega_2, \cdots, \omega_M\} \tag{3-1-1}$$

其中，ω_k 表示一个模式类；同时可定义**样本特征向量**(或称为待识别模式)为

$$\boldsymbol{X}_i = [x_{i1}, x_{i2}, \cdots, x_{id}]^{\mathrm{T}} \in X \subseteq \mathbb{R}^d, \quad i = 1, 2, \cdots, N \tag{3-1-2}$$

其中 $\boldsymbol{X}$ 表示样本特征向量空间，上标 T 表示转置；N 为样本点数；x_{ij} 是一个样本特征，d 为样本特征数；所谓**统计模式识别方法**就是根据属性或特征来定义某个距离函数以判别样本 $\boldsymbol{X}_i$ 属于哪个 ω_k。或者说，把有特征相似性的样本在模式空间中划分为互相接近的若干模式类，并以特征进行“聚类”，从而达到模式分类的目的。

统计模式识别的主要方法有：判别函数法、k 近邻分类法、非线性映射法、特征分析法、主因子分析法等。在统计模式识别中，Bayes 决策规则从理论上解决了最优分类器的设计问题，但其实施却必须首先解决更困难的概率密度估计问题。反向传播(BP)神经网络直接从观测数据(训练样本)学习，是更简便有效的方法，因而获得了广泛的应用。但它是一种启发式技术，缺乏坚实的理论基础。统计推断理论研究所取得的突破性成果导致现代统计学习理论——Vapnik Chervonenkis(VC)理论的建立，该理论不仅在严格的数学基础上圆满地回答了人工神经网络中出现的理论问题，而且导出了一种新的学习方法——支持向量机方法。

所谓智能学习就是从样本获得分类的一类数学方法，这些方法具有自学习的智能。

20 世纪 70 年代，波兰学者 Pawlak 等提出了**粗糙集(rough sets)**理论，此后，许多科学家在粗糙集的理论和应用方面做了大量的研究工作。至今，粗糙集理论已经成为智能学习的一类标准方法。粗糙集理论是处理模糊和不确定性的一个新的数学工具，用于数据分类分析。粗糙集理论分析的主要目的在于由获取的数据综合出近似的概念。本章 3.2 节将详细讨论粗糙集理论及应用。然而，Pawlak 的粗糙集理论是基于等价关系建立起来的，也就是说分类是严格的，但实际问题往往并不能进行严格分类，于是就引入**广义粗糙集(generalized rough sets)**的概念，并在此基础上建立容差关系的分类方法。20 世纪 70 年代初，Dempster 和 Shafer 建立的一套数学理论，称为 **D-S 证据理论(evidence theory)**，这一理论的核心是超越了概率统计推断的理论框架，可以适应于专家系统、人工智能、模式识别和系统决策等领域的实际问题。而且这一理论的发展很快成为智能学习和多源信息融合的重要组成部分。本章 3.3 节将对 D-S 证据理论进行详细研究。

G. Matheron 于 1975 年出版了《随机集与积分几何学》一书，从而提出了所谓**随机集(random sets)**理论，现在已经成为智能学习的又一重要研究方向。随机集处理的是随机集合问题，集值的引入带来许多令人意想不到的奇特结果，有可能发展成为一种更有力

的工具。因为它对异类多源信息融合有特别的意义，本章 3.4 节将专门讨论随机集理论。进而，我们还将介绍随机有限集理论的基本知识。

20 世纪 90 年代中期，统计学习理论和**支持向量机**（**support vector machine**）算法的深入研究和成功应用引起了广泛的关注。支持向量机算法具有比较坚实的理论基础和良好的推广能力，并在手写数字识别和文本分类等方面取得了良好的应用效果。特别是利用满足 Mercer 条件的核函数实现非线性分类，对于模式识别而言具有划时代的意义。

Bayes 网络是另外一种用于智能推断的工具，本章 3.7 节将进行简单的讨论。

模式识别发展至今，人们已经建立了各种针对具体应用问题的模型和方法，但仍然没有得到对所有模式识别问题都适用的统一模型和统一方法。我们现在拥有的只是一个工具箱，所要完成的工作就是结合具体的问题，把统计模式识别或句法模式识别结合起来，把模式识别方法与人工智能中的启发式搜索方法结合起来，把模式识别方法与支持向量机的机器学习方法结合起来，把人工神经元网络与各种已有技术以及人工智能中的专家系统、不确定推理方法结合起来，深入掌握各种工具的效能和应有的可能性，互相取长补短，不断开创模式识别应用的新局面。

3.2 粗糙集理论基础

粗糙集理论由 Zdzislaw Pawlak 在 20 世纪 80 年代早期提出[3-4]，用于数据分类分析。这些数据可以来自量测，也可以来自人类专家。粗糙集理论分析的主要目的在于由获取的数据综合出近似的概念。

3.2.1 信息系统的一般概念

定义 3.2.1 一个数据表$\mathcal{A}=(U,A)$称为是一个**信息系统**（**information system**），其中$U=\{x_1,x_2,\cdots,x_n\}$称为**对象集**（**set of objects**），$A=\{a_1,a_2,\cdots,a_m\}$称为**属性集**（**set of attributes**），都是非空有限集合；而$a:U\to V_a$，$\forall a\in A$描述了对应关系，其中集合V_a是属性a的**值集**（**value set**）。

例题 3.2.1 设有一个信息系统，示于表 3-2-1，其中有七个对象，两个属性。

表 3-2-1 信息系统$\mathcal{A}=(U,A)$

案例(U)	年龄段(a_1)	近三年患病次数(a_2)
x_1	16～30	50
x_2	16～30	0
x_3	31～45	1～25
x_4	31～45	1～25
x_5	46～60	26～49
x_6	16～30	26～49
x_7	46～60	26～49

注意，表中案例 3 和 4，以及案例 5 和 7 具有完全相同的条件值。在利用这些属性的

前提下，这些案例就是成对难识别的案例。 □

定义 3.2.2 一个**决策系统**(**decision system**)就是形式如$\mathcal{A}=(U,A\cup\{d\})$的信息系统，其中$d\notin A$称为**决策属性**(**decision attribute**)，相应地把A中的元素称为**条件属性**(**conditional attribute**)，或简称为**条件**(**condition**)。

决策属性通常取值为二值，即“是”或“否”(1或0)，但也可以取多值。

例题 3.2.2 针对例题3.2.1，增加决策属性，见表3-2-2，其中决策是决定是否复查，所以决策属性取二值。

表 3-2-2 决策系统$\mathcal{A}=(U,A\cup\{d\})$

案例(U)	年龄段(a_1)	近三年患病次数(a_2)	复查(d)
x_1	16～30	50	是
x_2	16～30	0	否
x_3	31～45	1～25	否
x_4	31～45	1～25	是
x_5	46～60	26～49	否
x_6	16～30	26～49	是
x_7	46～60	26～49	否

请注意，表中案例x_3和x_4，以及案例x_5和x_7仍然具有完全相同的条件值。但第一对取相反的决策值，而第二对取相同的决策值。 □

由决策表就可以综合出如下规则：“如果年龄在16～30年龄段，近3年来患病50次，身体复查是必需的。”

在许多应用中，有一种分类结果是已知的，就可以把这种后验知识用决策属性来表示，把这一过程就说成是**有监督学习过程**。

3.2.2 决策系统的不可分辨性

信息系统的知识发现问题本质上是按照属性特征给对象进行分类的问题。为此引入有关关系的概念[1]。

定义 3.2.3 设有对象集U，其笛卡儿积的子集$R\subseteq U\times U$就称为U上的一个**二元关系R**。进而，如果这个二元关系还具有如下特性：

(1) 自返性，如果$\forall x\in U\Rightarrow(x,x,)\in R$；

(2) 对称性，如果$(x,y)\in R\Rightarrow(y,x)\in R$；

(3) 传递性，如果$(x,y),(y,z)\in R\Rightarrow(x,z)\in R$，则称其为**等价关系**。

定义 3.2.4 设R是定义在对象集U上的一个等价关系，$x\in U$的一个**等价类**定义为

$$R(x)\triangleq\{y\in U:(x,y)\in R\} \tag{3-2-1}$$

定义 3.2.5 对象集U的某些子集形成的集类$\mathbb{U}=\{U_i\subseteq U:i\in I\}$，如果

(1) $\bigcup\limits_{i\in I}U_i=U$，$I$是指标集合；

(2) 对于任意$i\neq j\Rightarrow U_i\cap U_j=\varnothing$，则称$\mathbb{U}$为$U$的一个**划分**。

引理 1 设 R 是定义在对象集 U 上的一个等价关系，$R(x)$ 是 $x\in U$ 的等价类，如果 $y\in R(x)$，则必然有 $R(y)=R(x)$。

证明 见文献[1]。 ■

引理 2 设 R 是定义在对象集 U 上的一个等价关系，对于 $\forall x,y\in U$，如果 $R(y)\cap R(x)\neq\varnothing$，则必然有 $R(y)=R(x)$。

证明 见文献[1]。 ■

引理 3 设 R 是定义在对象集 U 上的一个等价关系，$R(x)$ 是 $x\in U$ 的等价类，则

$$\bigcup_{x\in U}R(x)=U \tag{3-2-2}$$

证明 见文献[1]。 ■

定理 3.2.1 定义在对象集 U 上的一个等价关系 R，确定了 U 对等价类的一个划分；反之，U 的任一划分也定义了 U 上的一个等价关系。

证明 见文献[1]。 ■

定义 3.2.6 对象集 U 按等价关系 R 形成的等价类所构成的集类称为 U 关于 R 的**商集**，记为 U/R。

根据定理 3.2.1，商集 U/R 实际上就是 U 按 R 的一个划分。

例题 3.2.3 设 $U=\mathbb{Z}^+$ 表示全体非负整数集合，其上定义了等价关系 R 如下

$$(x,y)\in R\Rightarrow x-y \text{ 能被 3 整除},\quad \forall x,y\in U$$

容易验证 R 是等价关系：① 自返性，$\forall x\in U\Rightarrow x-x=0\Rightarrow(x,x)\in R$；② 对称性，$\forall x,y\in U,(x,y)\in R\Rightarrow x-y$ 能被 3 整除$\Rightarrow y-x$ 能被 3 整除，即$(y,x)\in R$；③传递性，$(x,y),(y,z)\in R\Rightarrow x-y,y-z$ 能被 3 整除$\Rightarrow x-z=(x-y)+(y-z)$能被 3 整除，即$(x,z)\in R$。于是可以得到三个等价类：

$$R(1)=\{1,4,7,10,\cdots\},\quad R(2)=\{2,5,8,11,\cdots\},\quad R(3)=\{0,3,6,9,\cdots\}$$

而$\{R(1),R(2),R(3)\}$就是 U 按 R 的一个划分，即 $U/R=\{R(1),R(2),R(3)\}$。 □

一个决策系统(即决策表)表达了有关模型的所有知识。这个决策表的一部分可能没有必要那样大，因为至少按两种方式衡量有多余，即相同的或不可识别的对象可能被表示了几次，或某些属性可能是多余的。

定义 3.2.7 设 $\mathcal{A}=(U,A)$ 是一个信息系统，与任意 $B\subseteq A$ 相联系的等价关系定义为

$$\mathrm{IND}_{\mathcal{A}}(B)\triangleq\{(x,x')\in U\times U:\ \forall a\in B,a(x)=a(x')\} \tag{3-2-3}$$

称其为 ***B*-不可分辨关系（*B*-indiscernibility relation）**。如果$(x,x')\in\mathrm{IND}_{\mathcal{A}}(B)$，那么按 B 的属性，x 和 x'是不可分辨的。***B*-不可分辨关系的等价类**记为$[x]_B$。

如果不引起混淆的话，B-不可分辨关系的下标$\mathcal{A}$可以略去。

例题 3.2.4 对于表 3-2-1，条件属性 A 的非空子集有 $B_1=\{a_1\}=\{$年龄段$\}$、$B_2=\{a_2\}=\{$近三年患病次数$\}$，以及 $B_3=\{a_1,a_2\}=\{$年龄段，近三年患病次数$\}$，从而有如下三个 B-不可分辨关系

$$\mathrm{IND}(B_1)=\{\{x_1,x_2,x_6\},\{x_3,x_4\},\{x_5,x_7\}\}$$
$$\mathrm{IND}(B_2)=\{\{x_1\},\{x_2\},\{x_3,x_4\},\{x_5,x_6,x_7\}\}$$
$$\mathrm{IND}(B_3)=\{\{x_1\},\{x_2\},\{x_3,x_4\},\{x_5,x_7\},\{x_6\}\}$$

而 $[x_1]_{B_1}=\{x_1,x_2,x_6\}$,$[x_1]_{B_2}=\{x_1\}$;$[x_6]_{B_2}=\{x_5,x_6,x_7\}$,$[x_6]_{B_3}=\{x_6\}$,等等。

□

3.2.3 集合近似

定义 3.2.8 设$\mathcal{A}=(U,A)$是一个信息系统,$B\subseteq A$,$X\subseteq U$,定义 U 的子集合

$$\underline{B}(X)\triangleq\{x\in U:[x]_B\subseteq X\} \tag{3-2-4}$$

称为集合 X 的***B*-下近似(*B*-lower approximation)**;而 U 的另一子集合

$$\overline{B}(X)\triangleq\{x\in U:[x]_B\cap X\neq\varnothing\} \tag{3-2-5}$$

称为集合 X 的***B*-上近似(*B*-upper approximation)**。

$\underline{B}(X)$中的对象是基于 B 中的知识确定性地分类成 X 中的元;而 $\overline{B}(X)$中的对象是基于 B 中的知识仅仅分类成 X 中可能的元。

定义 3.2.9 集合

$$\mathrm{BN}_B(X)\triangleq\overline{B}(X)-\underline{B}(X) \tag{3-2-6}$$

称为 X 的***B*-边界区域(*B*-boundary region)**,是由不能基于 B 中的知识确定性地分类成 X 中元的对象组成。而

$$\mathrm{OT}_B(X)\triangleq U-\overline{B}(X) \tag{3-2-7}$$

称为 X 的***B*-外部区域(*B*-outside region)**,是由基于 B 中的知识确定性地分类成不属于 X 中元素的对象组成。

如果对象集合 U 一个子集 X 的 B-边界区域非空,则称其为**粗糙集(rough set)**,或简称为**粗集**;否则,如果其 B-边界区域是空集,则称其为**明晰集(crisp set)**。

例题 3.2.5 最常见的情况是按照条件属性综合决策类。在表 3-2-2 中,设结果集是

$$W=\{x\in U:d(x)=\{是\}\}=\{x_1,x_4,x_6\}\subset U$$

于是得到:$\underline{A}(W)=\{x_1,x_6\}$是集合 W 的 A-下近似,表示其等价类在 W 中的对象集合;$\overline{A}(W)=\{x_1,x_3,x_4,x_6\}$是集合 W 的 A-上近似,表示其等价类与 W 相交非空的对象集合;$\mathrm{BN}_A(W)=\{x_3,x_4\}$是 W 的 A-边界区域,是由不能基于 A 中的知识确定性地分类成 W 中元素的对象组成;$\mathrm{OT}_A(W)=\{x_2,x_5,x_7\}$,$W$ 的 A-外部区域,是由基于 A 中的知识确定性地分类成不属于 W 中元素的对象组成。所以 W 是一个粗集,因为它的边界区域非空。 □

集合近似图形表示如图 3-2-1 所示。

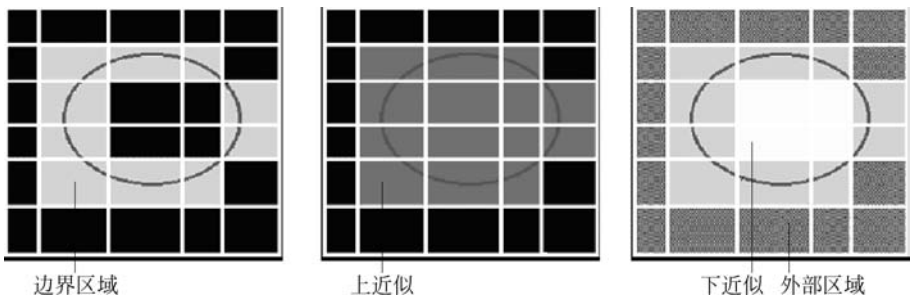


图 3-2-1 集合近似的图形表示

定理 3.2.2 设$\mathcal{A}=(U,A)$是一个信息系统，$B\subseteq A$，$X\subseteq U$，则有：

(1) $\underline{B}(X)\subseteq X\subseteq\overline{B}(X)$ (3-2-8)

(2) $\underline{B}(\varnothing)=\overline{B}(\varnothing)=\varnothing$，　$\underline{B}(U)=\overline{B}(U)=U$ (3-2-9)

(3) $\overline{B}(X\cup Y)=\overline{B}(X)\cup\overline{B}(Y)$，　$\underline{B}(X\cap Y)=\underline{B}(X)\cap\underline{B}(Y)$ (3-2-10)

(4) $X\subseteq Y\Rightarrow\underline{B}(X)\subseteq\underline{B}(Y)$且$\overline{B}(X)\subseteq\overline{B}(Y)$ (3-2-11)

(5) $\underline{B}(X\cup Y)\supseteq\underline{B}(X)\cup\underline{B}(Y)$，　$\overline{B}(X\cap Y)\subseteq\overline{B}(X)\cap\overline{B}(Y)$ (3-2-12)

(6) $\underline{B}(U-X)=U-\overline{B}(X)$，　$\overline{B}(U-X)=U-\underline{B}(X)$ (3-2-13)

(7) $\underline{B}(\underline{B}(X))=\overline{B}(\underline{B}(X))=\underline{B}(X)$，　$\overline{B}(\overline{B}(X))=\underline{B}(\overline{B}(X))=\overline{B}(X)$ (3-2-14)

证明　我们只证明式(3-2-8)和式(3-2-10)，其他可以类似证明。

(1) $x\in\underline{B}(X)\Rightarrow[x]_B\subseteq X\Rightarrow x\in X\Rightarrow[x]_B\cap X\neq\varnothing\Rightarrow x\in\overline{B}(X)$

(3) $x\in\overline{B}(X\cup Y)\Rightarrow\exists x\in U,[x]_B\cap(X\cup Y)\neq\varnothing\Rightarrow[x]_B\cap X\neq\varnothing$或$[x]_B\cap Y\neq\varnothing$

$\Rightarrow x\in\overline{B}(X)$　或　$x\in\overline{B}(Y)\Rightarrow x\in\overline{B}(X)\cup\overline{B}(Y)$

$x\in\underline{B}(X\cap Y)\Rightarrow\exists x\in U,[x]_B\subseteq(X\cap Y)\Rightarrow[x]_B\subseteq X$ 且$[x]_B\subseteq Y\Rightarrow x\in\underline{B}(X)$，且$x\in\underline{B}(Y)\Rightarrow x\in\underline{B}(X)\cap\underline{B}(Y)$ ■

显然，集合的下近似和上近似分别是由不可分辨关系产生的拓扑上集合的内部和闭包。图 3-2-2 给出了表 3-2-2 决策系统集合近似特性的图形表示。

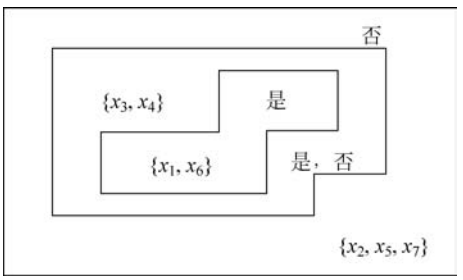


图 3-2-2　表 3-2-2 决策系统集合近似特性的图形表示

定义 3.2.10　可以定义如下四种粗集：

(1) X 是**粗糙 *B*-可定义的**（**roughly *B*-definable**），当且仅当$\underline{B}(X)\neq\varnothing$且$\overline{B}(X)\neq U$；

(2) X 是**内部 *B*-不可定义的**（**internally *B*-undefinable**），当且仅当$\underline{B}(X)=\varnothing$且$\overline{B}(X)\neq U$；

(3) X 是**外部 *B*-不可定义的**（**externally *B*-undefinable**），当且仅当$\underline{B}(X)\neq\varnothing$且$\overline{B}(X)=U$；

(4) X 是**完全 *B*-不可定义的**（**totally *B*-undefinable**），当且仅当$\underline{B}(X)=\varnothing$且$\overline{B}(X)=U$。

X 是粗糙 B-可定义的集合，就意味着借助于集合 B，就能确定 U 中的某些元属于 X，以及 U 中的另外一些元属于 $U-X$。

X 是内部 B-不可定义的集合，就意味着利用集合 B，我们就能确定 U 中的某些元属于 $U-X$，但不能确定 U 中的任何元是否属于 X。

X 是外部 B-不可定义的集合，就意味着利用集合 B，我们就能确定 U 中的某些元属于 X，但不能确定 U 中的任何元是否属于 $U-X$。

X 是完全 B-不可定义的集合，就意味着利用集合 B，我们既不能确定 U 中的任何元是否属于 X，也不能确定 U 中的任何元是否属于 $U-X$。

定义 3.2.11　粗集也可以用如下系数进行数值表示：

$$\alpha_B(X)=\frac{|\underline{B}(X)|}{|\overline{B}(X)|} \tag{3-2-15}$$

$\alpha_B(X)$称为**近似精度**(**accuracy of approximation**),其中$|X|$表示集合X的势(当其为有限集时,就是所包含元的个数);而

$$1-\alpha_B(X) \tag{3-2-16}$$

称为**近似粗糙度**(**roughness of approximation**)。

显然,$0\leqslant\alpha_B(X)\leqslant 1$。如果$\alpha_B(X)=1$,$X$关于属性集合$B$是明晰集,或称关于$B$是**精确的**(**precise**);否则,$\alpha_B(X)<1$,X关于属性集合B是粗集,或称关于B是**含糊的**(**vague**)。

3.2.4 属性约简

前一节我们研究了等价类的数据约简问题,即利用给定的属性,在对象集中建立不可分辨的等价类,即完成聚类。我们面临的另一类问题是尽可能使条件属性集变小,并保留原有属性集的分类能力,这就是**属性约简**(**attribute reduction**)的问题。然而,计算等价类是容易的,求得最小约简子集(即在所有约简子集中具有最小集合势的约简子集)却是一个NP难题。事实上,这正是把粗集理论应用于实际的一个瓶颈问题。已经发展了一些启发式算法,除了属性集合很大的情况外,可以在能接受的时间范围内计算充分多的约简子集。

例题 3.2.6 考虑表3-2-3的决策表,$\mathcal{A}'=(U,\{文凭,经历,外语,介绍人意见\}\cup\{决策\})$,我们只考虑条件属性,即信息系统$\mathcal{A}=(U,\{文凭,经历,外语,介绍人意见\})$。为了简单起见,每个等价类只包含一个元。

表 3-2-3 招聘问题:一个不可约简的决策表

对象(U)	文凭(d)	经历(e)	外语(f)	介绍人意见(r)	决策(dc)
x_1	MBA	中	通过	很好	接收
x_2	MBA	低	通过	不明确	拒绝
x_3	MCE	低	通过	好	拒绝
x_4	MSc	高	通过	不明确	接收
x_5	MSc	中	通过	不明确	拒绝
x_6	MSc	高	通过	很好	接收
x_7	MBA	高	否	好	接收
x_8	MCE	低	否	很好	拒绝

注:MBA:工商硕士;MCE:土木工程硕士;MSc:硕士。

似乎有一个最小的属性集$\{e,r\}$,用它识别目标与利用全部属性集识别目标是一样的。不妨利用全部属性集和利用属性集$\{e,r\}$来建立不可分辨关系,二者结果是一样的。具有这种特性的最小属性集的结构马上就显现出来了。 □

定义 3.2.12 给定一个信息系统$\mathcal{A}=(U,A)$,它的一个**约简**(**reduct**)就是一个最小属性集$B\subseteq A$,使得$\mathrm{IND}_{\mathcal{A}}(B)=\mathrm{IND}_{\mathcal{A}}(A)$。

换句话说,一个约简集就是属性集A中能保留对总对象集合划分特性的最小属性集合。因此,约简集的分类能力与总属性集的分类能力相同。

定义 3.2.13 给定一个具有n个对象的信息系统$\mathcal{A}=(U,A)$,它的**可分辨矩阵**

（**discernibility matrix**）就是一个 $n\times n$ 阵 $\boldsymbol{C}=\{c_{ij}\}$，且

$$c_{ij}=\{a\in A: a(x_i)\neq a(x_j)\},\quad i,j=1,2,\cdots,n \tag{3-2-17}$$

其中每个元都由一组属性构成，取决于对象 x_i 与 x_j 属性的不同。

定义 3.2.14 给定信息系统 $\mathcal{A}=(U,A)$，它的**可分辨函数** $f_{\mathcal{A}}$ 定义为 m 个 Boolean 变量 $a_1^*,a_2^*,\cdots,a_m^*$（相应于属性 $a_1,a_2,\cdots,a_m$）的 Boolean 函数，定义为

$$f_{\mathcal{A}}(a_1^*,a_2^*,\cdots,a_m^*)=\wedge\{\vee c_{ij}^*: 1\leqslant j\leqslant i\leqslant n, c_{ij}\neq\varnothing\} \tag{3-2-18}$$

其中 $c_{ij}^*=\{a^*: a\in c_{ij}\}$；而 $\wedge$ 表示逻辑“与”运算，$\vee$ 表示逻辑“或”运算。

注 $f_{\mathcal{A}}$ 的所有**主蕴涵**（**prime implicant**）构成的集合确定了 $\mathcal{A}$ 的所有约简构成的集合。所谓 Boolean 函数 f 的一个蕴涵就是字符（变量或者其“非”）间的某种关联，使得在变量的任意值 v 条件下字符取值为“真”时，函数 f 在 v 条件下的取值也为“真”。主蕴涵就是最小的蕴涵。我们只对单调 Boolean 函数的蕴涵感兴趣，因为这些函数不是由“非”来确定。

例题 3.2.7 仍考虑表 3-2-3 的决策表所定义的信息系统 $\mathcal{A}=(U,A)$，它的可分辨矩阵如表 3-2-4 所示。

表 3-2-4 可分辨矩阵 $\boldsymbol{C}=\{c_{ij}\}$

矩阵	$[x_1]$	$[x_2]$	$[x_3]$	$[x_4]$	$[x_5]$	$[x_6]$	$[x_7]$	$[x_8]$
$[x_1]$	$\varnothing$							
$[x_2]$	e,r	$\varnothing$						
$[x_3]$	d,e,r	d,r	$\varnothing$					
$[x_4]$	d,e,r	d,e	d,e,r	$\varnothing$				
$[x_5]$	d,r	d,e	d,e,r	e	$\varnothing$			
$[x_6]$	d,e	d,e,r	d,e,r	r	e,r	$\varnothing$		
$[x_7]$	e,f,r	e,f,r	d,e,f	d,f,r	d,e,f,r	d,f,r	$\varnothing$	
$[x_8]$	d,e,f	d,f,r	f,r	d,e,f,r	d,e,f,r	d,e,f	d,e,r	$\varnothing$

而可分辨函数是

$$\begin{aligned}f_{\mathcal{A}}(d,e,f,r)=&(e\vee r)(d\vee e\vee r)(d\vee e\vee r)(d\vee r)(d\vee e)(e\vee f\vee r)\\&(d\vee e\vee f)(d\vee r)(d\vee e)(d\vee e)(d\vee e\vee r)(e\vee f\vee r)\\&(d\vee f\vee r)(d\vee e\vee r)(d\vee e\vee r)(d\vee e\vee r)(d\vee e\vee f)\\&(f\vee r)(e)(r)(d\vee f\vee r)(d\vee e\vee f\vee r)\\&(e\vee r)(d\vee e\vee f\vee r)(d\vee e\vee f\vee r)\\&(d\vee f\vee r)(d\vee e\vee f)\\&(d\vee e\vee r)\end{aligned}$$

其中每个字符相应于每个属性，而每个括号内就是 Boolean 表达式的联结；而“与”运算符号均省略。经过函数的简化计算，结果就是 $f_{\mathcal{A}}(d,e,f,r)=er$（即 $e\wedge r$）。这与前述直观的约简结果是一致的。

注意，可分辨函数中的每一行相应于可分辨矩阵的每一列。而可分辨矩阵是一个对称阵，对角元为空集。于是，倒数第二行就意味着第六个对象（更确切地说是第六个等价

类),可以根据文凭(d)、外语(f)和介绍人意见(r)中的任意一个属性与第七个对象进行识别;可以根据文凭(d)、经历(e)和外语(f)中的任意一个属性与第八个对象进行识别。□

定义 3.2.15 如果在构造 Boolean 函数时,把 Boolean 变量的联结限制在可分辨矩阵中的 k 列,而不是所有的列,得到的函数称为 ***k*-相对可分辨函数**(***k*-relative discernibility function**);而由 k-相对可分辨函数得到的主蕴涵集就确定了 $\mathcal{A}=(U,A)$ 的所有 ***k*-相对约简**(***k*-relative reducts**)集。

这样的约简展现了由其他对象识别 $x_k\in U$(更确切地说是$[x_k]\subseteq U$)所需要的最小信息量。

利用上述概念,有监督学习问题(即分类结果已知的问题)就是求取决策 d 的值,这个值应当赋予新的对象,而这个对象的描述借助于条件属性。我们经常要求用来定义对象的属性集达到最小,例如,表 3-2-3 中出现了两个最小属性集{经历(e),介绍人意见(r)}和{文凭(d),经历(e)},都唯一地定义了对象隶属的决策类。相应的可分辨函数是相对于决策的。现在给出这个概念的形式化描述。

定义 3.2.16 设决策系统 $\mathcal{A}=(U,A\cup\{d\})$ 给定,映像 $d(U)=\{k:d(x)=k,x\in U\}$ 的势称为 d 的**秩**(**rank**),记为 $r(d)$,此处假定决策 d 的值集为 $V_d=\{v_d^{(1)},v_d^{(2)},\cdots,v_d^{(r(d))}\}$。

相当多的决策系统中 d 的秩是 2,例如,表 3-2-3 中的决策属性中就是{是,否}或{接收,拒绝}。然而 d 的秩可以是任意自然数。例如,表 3-2-3 中的决策属性中可以是{接收,保留,拒绝},则 d 的秩就成为 3。

定义 3.2.17 决策集 d 确定了对象集 U 的一个划分

$$\mathrm{CLASS}_{\mathcal{A}}=\{X_{\mathcal{A}}^{(1)},X_{\mathcal{A}}^{(2)},\cdots,X_{\mathcal{A}}^{(r(d))}\}$$

其中等价类 $X_{\mathcal{A}}^{(k)}=\{x\in U:d(x)=v_d^{(k)}\},k=1,2,\cdots,r(d)$;这个划分被称为决策系统**按决策的对象分类**(**classification of objects determined by the decision**);相应地称 $X_{\mathcal{A}}^{(k)}$ 为 ***A*的第 *k* 个决策类**(***k*-th decision class**),且等价类 $X_{\mathcal{A}}(u)=\{x\in U:d(x)=d(u)\}$,$\forall u\in U$。

例题 3.2.8 考虑表 3-2-2 的决策表,有如下按决策的对象分类

$$\mathrm{CLASS}_{\mathcal{A}}=\{X_{\mathcal{A}}^{(是)},X_{\mathcal{A}}^{(否)}\},\quad X_{\mathcal{A}}^{(是)}=\{x_1,x_4,x_6\},\quad X_{\mathcal{A}}^{(否)}=\{x_2,x_3,x_5,x_7\}$$

再考虑表 3-2-3 的决策表所定义的决策系统,则有如下按决策的对象分类

$$\mathrm{CLASS}_{\mathcal{A}}=\{X_{\mathcal{A}}^{(收)},X_{\mathcal{A}}^{(拒)}\},\quad X_{\mathcal{A}}^{(收)}=\{x_1,x_4,x_6,x_7\},\quad X_{\mathcal{A}}^{(拒)}=\{x_2,x_3,x_5,x_8\}$$

□

定义 3.2.18 设$\{X_{\mathcal{A}}^{(1)},X_{\mathcal{A}}^{(2)},\cdots,X_{\mathcal{A}}^{(r(d))}\}$是决策系统$\mathcal{A}$的决策类,则集合

$$\underline{B}(X_{\mathcal{A}}^{(1)})\cup\underline{B}(X_{\mathcal{A}}^{(2)})\cup\cdots\cup\underline{B}(X_{\mathcal{A}}^{(r(d))}) \tag{3-2-19}$$

称为是***A*的 *B*-正区域**(***B*-positive region**),记为$\mathrm{POS}_B(d)$。

例题 3.2.9 考虑决策表 3-2-2 所定义的决策系统 $\mathcal{A}=(U,A\cup\{d\})$,则有如下结果:$\underline{A}(X_{\mathcal{A}}^{(是)})\cup\underline{A}(X_{\mathcal{A}}^{(否)})\neq U$,这是因为对于对象 x_3,x_4,决策并不唯一。再考虑决策表 3-2-3 所定义的决策系统 $\mathcal{A}=(U,A\cup\{d\})$,则有 $\underline{A}(X_{\mathcal{A}}^{(收)})\cup\underline{A}(X_{\mathcal{A}}^{(拒)})=U$,因为此时所有的决策都是唯一的。□

定义 3.2.19 决策系统的这一重要特性可以形式化描述如下：设$\mathcal{A}=(U,A\cup\{d\})$是一个决策系统，而$\mathcal{A}$中的**广义决策**(**generalized decision**)定义为函数$\partial_A: U\to P(V_d)$，而

$$\partial_A(x)=\{i:\ \exists x'\in U,\text{使得}(x,x')\in \mathrm{IND}(A),\text{且 } d(x)=i\} \tag{3-2-20}$$

决策系统$\mathcal{A}$称为是**相容的**(**确定性的**)，如果对任意$x\in U$，总有$|\partial_A(x)|=1$；否则，决策系统$\mathcal{A}$称为是**不相容的**(**不确定性的**)。

容易看出，一个决策系统$\mathcal{A}$是相容的，当且仅当$\mathrm{POS}_A(d)=U$。进而，对于任意一对非空集合$B,B'\subseteq A$，如果$\partial_B=\partial_{B'}$，则$\mathrm{POS}_B(d)=\mathrm{POS}_{B'}(d)$。

例题 3.2.10 考虑决策表 3-2-2 所定义的决策系统$\mathcal{A}=(U,A\cup\{d\})$，其$\boldsymbol{A}$-正区域是$U$的一个子集；而表 3-2-3 所定义的决策系统$\mathcal{A}=(U,A\cup\{d\})$，其$\boldsymbol{A}$-正区域是整个$U$。前者是非确定性的，而后者是确定性的。 □

前面我们已经引入了k-相对可分辨函数的概念，因为决策属性是如此重要，对此引入专门的定义是非常有用的。

定义 3.2.20 设$\mathcal{A}=(U,A\cup\{d\})$是一个相容的决策系统，而$\boldsymbol{M}(\mathcal{A})=\{c_{ij}\}$是其可分辨矩阵，构造一个新的矩阵

$$\boldsymbol{M}^d(\mathcal{A})=\{c_{ij}^d\} \tag{3-2-21}$$

并假定

$$c_{ij}^d=\begin{cases}\varnothing, & d(x_i)=d(x_j)\\ c_{ij}-\{d\}, & \text{其他}\end{cases} \tag{3-2-22}$$

矩阵$\boldsymbol{M}^d(\mathcal{A})$称为系统$\mathcal{A}$的**决策相对可分辨矩阵**(**decision-relative discernibility matrix**)。与由可分辨矩阵构造可分辨函数相类似，可以由决策相对可分辨矩阵构造**决策相对可分辨函数**(**decision-relative discernibility function**)$f_M^d(\mathcal{A})$。

有关文献已经证明，$f_M^d(\mathcal{A})$的主蕴涵定义了$\mathcal{A}$的所有**决策相对约简**(**decision-relative reducts**)的集合。

例题 3.2.11 仍考虑表 3-2-3 所示的决策，重排后如表 3-2-5 所示。该表用来说明如何构造$\mathcal{A}$的决策相对可分辨矩阵和决策相对可分辨函数。为方便起见，把所有接收的行重排在上部，而所有拒绝的行重排在下部。

表 3-2-5 重排的决策表

对象	文凭	经历	外语	介绍人意见	决策
x_1	MBA	中	通过	很好	接收
x_4	MSc	高	通过	不明确	接收
x_6	MSc	高	通过	很好	接收
x_7	MBA	高	否	好	接收
x_2	MBA	低	通过	不明确	拒绝
x_3	MCE	低	通过	好	拒绝
x_5	MSc	中	通过	不明确	拒绝
x_8	MCE	低	否	很好	拒绝

由这个决策相对可分辨矩阵得到的简化决策相对可分辨函数是$f_M^d(\mathcal{A})=ed\vee er$。

由决策相对可分辨矩阵的定义知，选择其中的一个列，如相应于$[x_1]$的列，简化并给出最小函数，以识别$[x_1]$所属的决策类。例如，第一列给出了 Boolean 函数是：$(e \vee r)(d \vee e \vee r)(d \vee r)(d \vee e \vee f)$，经化简后变为 $ed \vee rd \vee re \vee rf$。读者可以检验："如果介绍人意见是很好，外语是通过，则接收"，这就是 x_1 的情况；其实，对于别的对象，如 x_6 也是同样情况。

相应的决策相对可分辨矩阵示于表 3-2-6，这是一个对称矩阵，其中对角元素为空集，而所有决策相同的元素也是空集。

表 3-2-6　决策相对可分辨矩阵

矩阵	$[x_1]$	$[x_4]$	$[x_6]$	$[x_7]$	$[x_2]$	$[x_3]$	$[x_5]$	$[x_8]$
$[x_1]$	$\varnothing$							
$[x_4]$	$\varnothing$	$\varnothing$						
$[x_6]$	$\varnothing$	$\varnothing$	$\varnothing$					
$[x_7]$	$\varnothing$	$\varnothing$	$\varnothing$	$\varnothing$				
$[x_2]$	e,r	d,e	d,e,r	e,f,r	$\varnothing$			
$[x_3]$	d,e,r	d,e,r	d,e,r	d,e,f	$\varnothing$	$\varnothing$		
$[x_5]$	d,r	e	e,r	d,e,f,r	$\varnothing$	$\varnothing$	$\varnothing$	
$[x_8]$	d,e,f	d,e,f,r	d,e,f	d,e,r	$\varnothing$	$\varnothing$	$\varnothing$	$\varnothing$

□

3.2.5　粗糙隶属度

定义 3.2.21　一个具有特征函数的集合称为**明晰集(crisp set)**，即 U 是对象集合(或论域)，$S \subseteq U$ 是子集，$\chi_S: U \to \{0,1\}$ 是**特征函数(characteristic function)**，则

$$\chi_S(x) = \begin{cases} 1, & x \in S \\ 0, & x \notin S \end{cases} \tag{3-2-23}$$

模糊集(fuzzy set)的概念是用**隶属度函数(membership function)**代替了特征函数，即 $m_S: U \to [0,1]$ 是隶属度函数，而

$$m_S(x) = \alpha, \quad 0 \leqslant \alpha \leqslant 1, \quad \forall x \in U \tag{3-2-24}$$

明晰集用于正规地表征一个概念，具有清晰的边界，因此并不反映隶属的不确定性。模糊集是对明晰集的推广，这一概念已经成功地得到应用。

模糊集合可以进行明晰特征化。

定义 3.2.22　设 $S \subseteq U$ 的支持集定义为

$$\text{support}_U(S) = \{x \in U: m_S(x) > 0\} \tag{3-2-25}$$

又设 $A, B \subseteq U$ 是模糊集合，其包含关系定义为

$$A \subseteq B \quad \text{iff} \quad m_A(x) \leqslant m_B(x), \quad \forall x \in U \tag{3-2-26}$$

(式中 iff 表示"当且仅当")而模糊集合的其他运算定义为

$$\text{并运算：} m_{A \cup B}(x) = \max\{m_A(x), m_B(x)\} \tag{3-2-27}$$

交运算：$m_{A\cap B}(x)=\min\{m_A(x),m_B(x)\}$ (3-2-28)

补运算：$m_{U-A}(x)=1-m_A(x)$ (3-2-29)

定义 3.2.23 模糊集合 X 和 Y 之间的**模糊关系**(**fuzzy relation**)R 定义为笛卡儿空间 $X\times Y$ 中的一个模糊集合，即

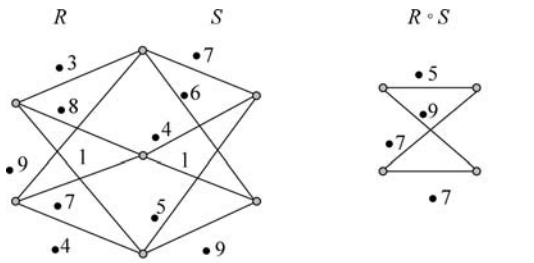
$$m_R: X\times Y\to[0,1] \tag{3-2-30}$$

是 x 和 y 在 R 中关系的隶属度。

定义 3.2.24 $X\times Y$ 中的模糊关系 R 和 $Y\times Z$ 中的模糊关系 S，合成为 $X\times Z$ 中的**复合模糊关系**(**composition fuzzy relation**)$R\circ S$，定义为

$$m_{R\circ S}(x,z)\triangleq\max_{y\in Y}\{\min[m_R(x,y),m_S(y,z)]\} \tag{3-2-31}$$

例题 3.2.12 考虑图 3-2-3 给出的模糊关系，表 3-2-7 描述了关系 R，表 3-2-8 描述了关系 S；图 3-2-3(a)表示两个模糊关系，图 3-2-3(b)表示复合的模糊关系。表 3-2-9～表 3-2-12 分别给出了复合计算过程。



(a) 模糊关系R,S　　(b) 复合模糊关系$R\circ S$

图 3-2-3　模糊关系的复合

表 3-2-7～表 3-2-12 给出了运算结果。

表 3-2-7　关系 *R*

m_R		y		
		1	2	3
x	1	0.3	0.8	1
	2	0.9	0.7	0.4

表 3-2-8　关系 *S*

m_S		z	
		1	2
y	1	0.7	0.6
	2	0.4	1
	3	0.5	0.9

表 3-2-9　复合计算过程 1

$\max\limits_y\{\min[m_R(x,y),m_S(y,z)]\},x=1,z=1$

y	$m_R(1,y)$	$m_S(y,1)$	min
1	0.3	0.7	0.3
2	0.8	0.4	0.4
3	1	0.5	0.5
max			0.5

表 3-2-10　复合计算过程 2

$\max\limits_y\{\min[m_R(x,y),m_S(y,z)]\},x=1,z=2$

y	$m_R(1,y)$	$m_S(y,2)$	min
1	0.3	0.6	0.3
2	0.8	1	0.8
3	1	0.9	0.9
max			0.9

表 3-2-11 复合计算过程 3

$\max\limits_{y}\{\min[m_R(x,y),m_S(y,z)]\},x=2,z=1$

y	$m_R(2,y)$	$m_S(y,1)$	min
1	0.9	0.7	0.7
2	0.7	0.4	0.4
3	0.4	0.5	0.4
max			0.7

表 3-2-12 复合计算过程 4

$\max\limits_{y}\{\min[m_R(x,y),m_S(y,z)]\},x=2,z=2$

y	$m_R(2,y)$	$m_S(y,2)$	min
1	0.9	0.6	0.6
2	0.7	1	0.7
3	0.4	0.9	0.4
max			0.7

□

定义 3.2.25 设$U=\{x_1,x_2,\cdots,x_n\}$是论域，p 是概率密度，A 是 U 中的模糊集(事件)，则**模糊事件的概率**(**probabilities of fuzzy events**)定义为

$$P(A)=\sum_{i=1}^{n}m_A(x_i)p(x_i) \tag{3-2-32}$$

其中 $m_A(x_i)$是隶属度。

定义 3.2.26 对论域 U 中的模糊集合 A 寻求一个单值来表示，就是**去模糊化**(**defuzzyfication**)，常用的方法有两个：

(1) 最大化法：

$$\hat{x}=\operatorname{argmax}_{x\in U}m_A(x) \tag{3-2-33}$$

(2) 求重心法：

$$\hat{x}=\frac{\sum_{i=1}^{n}x_i m_A(x_i)}{\sum_{i=1}^{n}m_A(x_i)} \tag{3-2-34}$$

定义 3.2.27 A 是论域 U 中的一个模糊集合，其**α -截取**(**alpha cuts**)定义为

$$A_\alpha\triangleq\{x\in U:m_A(x)\geqslant\alpha\} \tag{3-2-35}$$

其强**α -截取**(**strong alpha cuts**)定义为

$$A_\alpha\triangleq\{x\in U:m_A(x)>\alpha\} \tag{3-2-36}$$

A 的α-截取是明晰集。

定义 3.2.28 设$\mathcal{A}=(U,A)$是一个信息系统，$B\subseteq A$，$X\subseteq U$，**粗糙隶属函数**(**rough membership function**)定义为集合 X 与包含 x 的等价类$[x]_B$ 之间**相对交迭度**(**degree of relative overlap**)的一种定量化描述，即

$$\mu_X^B:U\to[0,1],\quad \text{且}\quad \mu_X^B=\frac{|[x]_B\cap X|}{|[x]_B|} \tag{3-2-37}$$

粗糙隶属度函数可以解释为给定 x 关于属性 B 的信息$u=\mathrm{Inf}_B(x)$的前提下，x 属于 X 的条件概率$P(x\in X|u)$的频度估计。

3.2.6 广义粗集

等价关系是 Pawlak 经典粗集模型的一个关键概念，但等价关系的要求对众多应用

来讲太过严格，从而大大限制了粗集理论的推广应用。因此，粗集理论研究的一个重要方向便是建立关于非等价关系的广义粗集[5-8]模型，为此，许多学者做了大量工作，必须对一般关系下的粗集模型进行扩充。

定义 3.2.29 设$(U,A\cup\{d\})$是一个决策系统，集合U上关于条件属性A的**容差关系**定义为

$$\tau_A^{\varepsilon}\triangleq\{(u,v)\in U\times U:\rho_A(u,v)\leqslant\varepsilon\} \tag{3-2-38}$$

其中$\rho_A:U\times U\to\mathbb{R}^+$是相对于条件属性集$A$的对象间距离，$\varepsilon$是一个阈值，则称$(U,A\cup\{d\},\tau_A^{\varepsilon})$为**广义粗集模型**。

类似地，考虑A的子集B，条件属性子集B上的容差关系定义为

$$\tau_B^{\varepsilon}=\{(u,v)\in U\times U:\rho_B(u,v)\leqslant\varepsilon\} \tag{3-2-39}$$

其中距离函数定义为$\rho_B(u,v)\triangleq\max_{a\in B}\{\mathrm{dist}(a(u),a(v))\}$，其中$\rho_B:U\times U\to\mathbb{R}^+$是相对于条件属性集$B$的对象间距离；而$\mathrm{dist}:V_a\times V_a\to\mathbb{R}^+$是相对属性$a\in B$的值集上的距离函数。

注 广义粗集不是粗集，因为容差关系不是等价关系，它虽然满足二元关系的自返性和对称性，但不满足传递性！也就是说，$(u,v),(v,w)\in\tau_A^{\varepsilon}$，但不能推出$(u,w)\in\tau_A^{\varepsilon}$。

类似地，求得一个信息系统的**广义属性约简**就是寻找一个最小属性集合，使得它对所有对象的广义分类能力等同于原来属性集合的广义分类能力。

这仍然是一个难以求解的问题。

3.3 证据理论基础

证据理论是由 Dempster 和 Shafer 于 20 世纪 60 年代末和 70 年代初建立的一套数学理论，是对概率论的进一步扩充，适合于专家系统、人工智能、模式识别和系统决策等领域的实际问题。

3.3.1 概述

例题 3.3.1 设某甲有两个硬币，一个是正常的正反面硬币，另一个是两面都是正面的错币。他投掷硬币 10 次，每次都出现的是“正面”，问下一次投掷出现正面的概率是多大？我们可以按照古典概率来计算：

P(正面)=0.5——如果某甲持正常硬币；

P(正面)=1.0——如果某甲持错误硬币。

我们缺少的正是某甲持哪个硬币的“证据”。

首先我们考虑用 Bayes 方法来解决。假设某甲持错误硬币的先验概率是α，则N次投掷出现正面后持错误硬币的后验概率是

$$P(\text{错误硬币}\mid N\text{次出现正面})=\frac{2^N\alpha}{2^N\alpha+(1-\alpha)}$$

显然随着N的增大，这个后验概率很快趋于 1。但是，这个先验概率如何得到？是否有

其他方法无须给出先验概率，而凭直觉能获得这个硬币是“错币”的证据？ □

以前处理类似问题只能利用概率论中事件概率的框架，而现在某些感兴趣的因素却不能用概率的方法来处理。

事实上，有关证据问题在哲学文献里已经做过很深入的研究，而证据在本质上就是基于观测对不同的假设赋予权值的一种方法。根据有关证据的定义，我们简单地给出如下解释：

(1) 能够处理任意数量的假设；

(2) 能够把证据的权值解释为一种函数，而这个函数把假设的先验概率空间，映射到基于观测的假设后验概率空间。

3.3.2 mass 函数、信度函数与似真度函数

定义 3.3.1 设$\mathcal{H}$表示某个有限集合，称为**假设空间**，这是一个完备的、元素间相互不交的空间；又假定$\mathcal{O}$表示**观测空间**，或称为试验结果集合。对于给定的假设 $h\in\mathcal{H}$，μ_h是观测空间$\mathcal{O}$上的概率，而**证据空间**定义为

$$\mathcal{E}\triangleq\{\mathcal{H},\mathcal{O},\mu_{h_1},\mu_{h_2},\cdots,\mu_{h_n}\} \tag{3-3-1}$$

其中 $n\in\mathbb{N}$ 是$\mathcal{H}$中假设的个数。

例题 3.3.2 在例题 3.3.1 中，$\mathcal{H}=\{h_1=\{$正常硬币$\},h_2=\{$错误硬币$\}\}$，而$\mathcal{O}=\{z:z=\{$观测 1，观测 2，…，观测 10$\}\}$。

设 $z_1=\{$观测 1$=\{$正面$\}$，观测 2$=\{$正面$\}$，…，观测 10$=\{$正面$\}\}$是一个具体的观测，则

$$\mu_{h_1}(z_1)=\frac{1}{2^{10}},\quad \mu_{h_2}(z_1)=1$$ □

给定假设 $h\in\mathcal{H}$，以及观测 $z\in\mathcal{O}$，在证据空间中权值为

$$w(z,h)=\frac{\mu_h(z)}{\mu_{h_1}(z)+\cdots+\mu_{h_n}(z)} \tag{3-3-2}$$

注意，不失一般性，考察 h_1 和 h_2 的情况，有 $w(z,h_1)/w(z,h_2)=\mu_{h_1}(z)/\mu_{h_2}(z)$。

所以，给定 $h\in\mathcal{H}$，观测 z 的权值就是其相对概率，而且 $w(z,h)\in[0,1]$；对于给定的 $z\in\mathcal{O}$，$w(z,\cdot)$就是$\mathcal{H}$上的一个概率测度。但是，这却不是一个频度或似然。例如，在例题 3.3.2 中，$w(z_1,h_2)=1/(1+1/2^{10})=2^{10}/(2^{10}+1)$。

定义 3.3.2 设$\mathcal{H}$表示某个有限集合，称为假设空间；又假定$\mathcal{P}(\mathcal{H})$表示$\mathcal{H}$的所有子集构成的集类(称为$\mathcal{H}$的幂集)，**映射 m**：$\mathcal{P}(\mathcal{H})\to[0,1]$称为一个**基本概率赋值**(**basic propability assignment**，**BPA**)或 **mass 函数**，如果

$$m(\varnothing)=0;\quad \sum_{A\subseteq\mathcal{H}}m(A)=1 \tag{3-3-3}$$

那么 mass 函数实际上就是对各种假设的评价权值。但是，基本概率赋值不是概率，因为不满足可列可加性。也就是说，如果 $A,B,C\subseteq\mathcal{H}$，且 $A=B\cup C$，$B\cap C=\varnothing$，但是 $m(A)=m(B)+m(C)$却不必成立。

例题 3.3.3 设有两个医生给同一病人诊断疾病，甲医生认为 0.9 的可能性是感冒，

0.1的可能性是说不清楚的病症；乙医生认为0.2的可能性不是感冒，0.8的可能性是说不清的病症。于是，假设空间是$\mathcal{H}=\{h,\bar{h}\}$，其中$h$表示诊断为感冒，$\bar{h}$表示诊断为不是感冒；$\mathcal{P}(\mathcal{H})=\{\varnothing,\{h\},\{\bar{h}\},\mathcal{H}\}$，其中$\varnothing$表示不可能事件："既是感冒，又不是感冒"，而$\mathcal{H}$表示事件："可能是感冒，又可能不是感冒"。从而可以构造所谓mass函数：

$m_1(h)=0.9$，表示甲医生认为是感冒的可能性；

$m_1(\mathcal{H})=0.1$，表示甲医生认为是说不清楚何种病症的可能性；

$m_2(\bar{h})=0.2$，表示乙医生认为不是感冒的可能性；

$m_2(\mathcal{H})=0.8$，表示乙医生认为是说不清楚何种病症的可能性。

问题是判定患者是感冒的可能性究竟有多大，或者判定这种可能性落在什么范围内。注意$\mathcal{H}=\{h\}\cup\{\bar{h}\}$，且$\{h\}\cap\{\bar{h}\}=\varnothing$，但

$$m_1(\mathcal{H})\neq m_1(\{h\})+m_1(\{\bar{h}\})$$ □

定义 3.3.3 设$\mathcal{H}$表示某个有限集合，$\mathcal{P}(\mathcal{H})$表示$\mathcal{H}$的所有子集构成的集类，映射$Bel:\mathcal{P}(\mathcal{H})\to[0,1]$称为**信度函数(belief function)**，如果：

(1) $Bel(\varnothing)=0$；$Bel(\mathcal{H})=1$； (3-3-4)

(2) 对$\mathcal{H}$中的任意子集$A_1,A_2,\cdots,A_n$有

$$Bel\Big(\bigcup_{i=1}^{n}A_i\Big)\geqslant\sum_{\substack{I\subseteq\{1,2,\cdots,n\}\\I\neq\varnothing}}(-1)^{|I|+1}Bel\Big(\bigcap_{i\in I}A_i\Big) \tag{3-3-5}$$

式中$|I|$表示集合I中元素的个数。

这说明信度函数表示对假设的信任程度估计的下限(悲观估计)。假定仅对$\mathcal{H}$中的任意两个子集A_1,A_2有

$$Bel(A_1\cup A_2)\geqslant Bel(A_1)+Bel(A_2)-Bel(A_1\cap A_2) \tag{3-3-6}$$

则称$Bel:\mathcal{P}(\mathcal{H})\to[0,1]$为**弱信度函数(weak belief function)**。

定义 3.3.4 设$\mathcal{H}$表示某个有限集合，$\mathcal{P}(\mathcal{H})$表示$\mathcal{H}$的所有子集构成的集类，映射$Pl:\mathcal{P}(\mathcal{H})\to[0,1]$称为**似真度函数(plausibility function)**，如果

(1) $Pl(\varnothing)=0$；$Pl(\mathcal{H})=1$； (3-3-7)

(2) 对$\mathcal{H}$中的任意子集$A_1,A_2,\cdots,A_n$有

$$Pl\Big(\bigcap_{i=1}^{n}A_i\Big)\leqslant\sum_{\substack{I\subseteq\{1,2,\cdots,n\}\\I\neq\varnothing}}(-1)^{|I|+1}Pl\Big(\bigcup_{i\in I}A_i\Big) \tag{3-3-8}$$

式中$|I|$表示集合I中元素的个数。

这说明似真度函数表示对假设的信任程度估计的上限(乐观估计)。假定仅对X中的任意两个子集A_1,A_2有

$$Pl(A_1\cap A_2)\leqslant Pl(A_1)+Pl(A_2)-Pl(A_1\cup A_2) \tag{3-3-9}$$

则称$Pl:\mathcal{P}(\mathcal{H})\to[0,1]$为**弱似真度函数(weak plausibility function)**。

例题 3.3.4 仍考虑诊断问题，假定先验知识告诉我们病人的疾病只有三种可能性h_1,h_2,h_3，构成基本假设空间$\mathcal{H}=\{h_1,h_2,h_3\}$，于是事件空间为

$$\mathcal{P}(\mathcal{H})=\{\varnothing,\{h_1\},\{h_2\},\{h_3\},\{h_1,h_2\},\{h_1,h_3\},\{h_2,h_3\},\{h_1,h_2,h_3\}\}$$

而三个医生分别按自己的诊断为上述事件空间给出了三个不同的概率P_1,P_2,P_3，它们

满足

$$\begin{cases} P_1(\{h_1\})=0.5, & P_1(\{h_2\})=0, & P_1(\{h_3\})=0.5 \\ P_2(\{h_1\})=0.5, & P_2(\{h_2\})=0.5, & P_2(\{h_3\})=0 \\ P_3(\{h_1\})=0, & P_3(\{h_2\})=0.5, & P_3(\{h_3\})=0.5 \end{cases}$$

按概率公式,可计算得到

$$\begin{cases} P_1(\{h_1,h_2\})=0.5, & P_1(\{h_1,h_3\})=1, & P_1(\{h_2,h_3\})=0.5 \\ P_2(\{h_1,h_2\})=1, & P_2(\{h_1,h_3\})=0.5, & P_2(\{h_2,h_3\})=0.5 \\ P_3(\{h_1,h_2\})=0.5, & P_3(\{h_1,h_3\})=0.5, & P_3(\{h_2,h_3\})=1 \end{cases}$$

$$\begin{cases} P_1(\varnothing)=P_2(\varnothing)=P_3(\varnothing)=0 \\ P_1(\mathcal{H})=P_2(\mathcal{H})=P_3(\mathcal{H})=1 \end{cases}$$

对于任意事件 $A\in\mathcal{P}(\mathcal{H})$,其**概率下界**和**概率上界**分别定义为

$$P_*(A)=\min_i\{P_i(A):i=1,2,3\},\quad P^*(A)=\max_i\{P_i(A):i=1,2,3\}$$

于是:

$$P_*(\{h_1\})=P_*(\{h_2\})=P_*(\{h_3\})=0$$

$$P_*(\{h_1,h_2\})=P_*(\{h_2,h_3\})=P_*(\{h_1,h_3\})=0.5$$

$$P_*(\mathcal{H})=1,P_*(\varnothing)=0$$

$$P^*(\{h_1\})=P^*(\{h_2\})=P^*(\{h_3\})=0.5$$

$$P^*(\{h_1,h_2\})=P^*(\{h_2,h_3\})=P^*(\{h_1,h_3\})=1$$

$$P^*(\mathcal{H})=1,P^*(\varnothing)=0$$

这就是概率下界和概率上界。显然,概率下界和概率上界具有如下性质:

(1) $P_*(\varnothing)=P^*(\varnothing)=0$; $P_*(\mathcal{H})=P^*(\mathcal{H})=1$

(2) $0\leqslant P_*(A)\leqslant P^*(A)\leqslant 1,\forall A\in\mathcal{P}(\mathcal{H})$

(3) $P^*(A)=1-P_*(A^c),\forall A\in P(\mathcal{H})$(其中 $A^c=\mathcal{H}-A$)

(4) $P_*(A)+P_*(B)\leqslant P_*(A\cup B)\leqslant P_*(A)+P^*(B)\leqslant P^*(A\cup B)\leqslant P^*(A)+P_*(B),\forall A,B\in\mathcal{P}(\mathcal{H}),A\cap B=\varnothing$

概率下界和概率上界虽然满足有界性,但一般不再满足可加性。

显然,因为 $P_*(A\cap B)\geqslant 0$,所以有

$$P_*(A\cup B)\geqslant P_*(A)+P_*(B)-P_*(A\cap B),\quad \forall A,B\in\mathcal{P}(\mathcal{H}) \tag{3-3-10}$$

从而 P_* 是弱信度函数。

同样地,因为

$$\begin{aligned} P^*(A\cap B)&=1-P_*(A^c\cup B^c)\leqslant 1-P_*(A^c)-P_*(B^c)+P_*(A^c\cap B^c) \\ &=1-P_*(A^c)+1-P_*(B^c)-1+P_*(A^c\cap B^c) \\ &=P^*(A)+P^*(B)-P^*(A\cup B) \end{aligned} \tag{3-3-11}$$

从而 P^* 是弱似真度函数。令 $A_1=\{h_1\},A_2=\{h_2\},A_3=\{h_3\}$,而

$$P^*(A_1\cap A_2\cap A_3)=P^*(\varnothing)=0$$

从而有

$$P^*(A_1)+P^*(A_2)+P^*(A_3)-P^*(A_1\cup A_2)-P^*(A_2\cup A_3)-P^*(A_1\cup A_3)+P^*(A_1\cup A_2\cup A_3)=0.5+0.5+0.5-1-1-1+1=-0.5$$

所以式(3-3-8)不成立，从而 P^* 不是似真度函数。 □

定理 3.3.1 设$\mathcal{H}$表示某个有限集合，Bel 和 Pl 分别表示$\mathcal{H}$上的信度函数和似真度函数，则有

$$Pl(A)=1-Bel(A^c) \tag{3-3-12}$$

$$Bel(A)=1-Pl(A^c) \tag{3-3-13}$$

证明 由定义及集合并交运算可证。 ■

定理 3.3.2 设 m、Bel 和 Pl 分别是$\mathcal{H}$上的 mass 函数、信度函数和似真度函数，则对任意 $A\in\mathcal{P}(\mathcal{H})$有

$$Bel(A)=\sum_{D\subseteq A}m(D) \tag{3-3-14}$$

$$Pl(A)=\sum_{D\cap A\neq\varnothing}m(D) \tag{3-3-15}$$

且 $Bel(A)\leqslant Pl(A)$。

证明 根据 mass 函数定义知，$Bel(\varnothing)=m(\varnothing)=0$，$Bel(\mathcal{H})=\sum_{A\subseteq\mathcal{H}}m(A)=1$，而且对任意 $A\in\mathcal{P}(\mathcal{H})$有 $Bel(A)\geqslant 0$，且根据信度函数的定义知

$$1=Bel(\mathcal{H})=Bel(A\cup A^c)\geqslant Bel(A)+Bel(A^c)$$

从而 $Bel(A)\leqslant 1-Bel(A^c)\leqslant 1$。对于 $D\subseteq\mathcal{H}$，设 $A_1,A_2,\cdots,A_n$ 是$\mathcal{H}$中的任意子集，记 $I(D)=\{1,2,\cdots,n\}\cap\{i:\ D\subseteq A_i\}$，显然 $I(D)\neq\varnothing\Leftrightarrow\exists\, i\in\{1,2,\cdots,n\}$，使得 $D\subseteq A_i$，而

$$D\subseteq\bigcap_{i\in I}A_i\Leftrightarrow I\subseteq I(D)$$

于是，

$$\begin{aligned}\sum_{\substack{I\subseteq\{1,2,\cdots,n\}\\ I\neq\varnothing}}(-1)^{|I|+1}Bel\Big(\bigcap_{i\in I}A_i\Big)&=\sum_{\substack{I\subseteq\{1,2,\cdots,n\}\\ I\neq\varnothing}}(-1)^{|I|+1}\Big[\sum_{I\subseteq I(D),I(D)\neq\varnothing}m(D)\Big]\\&=\sum_{I(D)\neq\varnothing}m(D)\Big[\sum_{I(D)\neq\varnothing,I\subseteq I(D)}(-1)^{|I|+1}\Big]=\sum_{I(D)\neq\varnothing}m(D)\\&\leqslant\sum\Big\{m(D):\ D\subseteq\bigcup_{i=1}^{n}A_i\Big\}=Bel\Big(\bigcup_{i=1}^{n}A_i\Big)\end{aligned}$$

所以 Bel 是信度函数。由 $Pl(A)=1-Bel(A^c)$即可知 Pl 是似真度函数。 ■

定义 3.3.5 设$\mathcal{H}$是有限集合，Bel 和 Pl 分别是定义在$\mathcal{P}(\mathcal{H})$上的信度函数和似真度函数，对于任意 $A\in\mathcal{P}(\mathcal{H})$，其**信度区间**(**belief interval**)定义为

$$[Bel(A),Pl(A)]\subseteq[0,1] \tag{3-3-16}$$

信度区间表示事件发生的下限估计到上限估计的可能范围。

定理 3.3.3 设 $z_1,z_2,\cdots,z_l\in\mathcal{O}$为 l 个互斥且完备的观测，即 $\mu(z_i)$表示 z_i 发生的概率，满足 $z_i\cap z_j=\varnothing$，$\forall\, i\neq j$ 且 $\sum_{i=1}^{l}\mu(z_i)=1$；对于每个 $z_i\in\mathcal{O}$，当 $m(\cdot\,|z_i)$，$Bel(\cdot\,|z_i)$，$Pl(\cdot\,|z_i)$分别是$\mathcal{H}$上的 mass 函数、信度函数和似真度函数时，则

$$m(A)=\sum_{i=1}^{l}m(A\mid z_i)\mu(z_i) \tag{3-3-17}$$

$$Bel(A)=\sum_{i=1}^{l}Bel(A\mid z_i)\mu(z_i) \tag{3-3-18}$$

$$Pl(A)=\sum_{i=1}^{l}Pl(A\mid z_i)\mu(z_i) \tag{3-3-19}$$

仍分别是$\mathcal{H}$上的 mass 函数、信度函数和似真度函数。

证明 容易证明 $m(\varnothing)=0$,而且

$$\sum_{A\subseteq\mathcal{H}}m(A)=\sum_{A\subseteq\mathcal{H}}\sum_{i=1}^{l}m(A\mid z_i)\mu(z_i)=\sum_{i=1}^{l}\mu(z_i)\sum_{A\subseteq\mathcal{H}}m(A\mid z_i)=1$$

从而 m 是 mass 函数。同理可证信度函数和似真度函数。 ■

例题 3.3.5 设某工厂要为某设备的故障 A 的发生率 $m(A)$做出判断。而根据统计,各指标分类等级发生的概率分别为 $\mu(z_1)=0.1,\mu(z_2)=0.15,\mu(z_3)=0.3,\mu(z_4)=0.25,\mu(z_5)=0.2$。已经知道在各检验指标 z 分类等级条件下故障 A 的发生率分别为:

(1) $m(A\mid z_1)=0.03$,z_1 表示严重超标;$m(A\mid z_2)=0.025$,z_2 表示超标;

(2) $m(A\mid z_3)=0.01$,z_3 表示轻微超标;$m(A\mid z_4)=0.005$,z_4 表示良好;

(3) $m(A\mid z_5)=0.001$,z_5 表示优良。

于是,故障 A 的发生率为

$$\begin{aligned}m(A)=&0.03\times0.1+0.025\times0.15+0.01\times0.3+\\&0.005\times0.25+0.001\times0.2=0.0112\end{aligned}$$ □

3.3.3 Dempster 公式

定理 3.3.4(Dempster 公式) 设 m_1,m_2 是$\mathcal{H}$上的两个 mass 函数,则

$$m(\varnothing)=0$$

$$m(A)=\frac{1}{N}\sum_{E\cap F=A}m_1(E)m_2(F),\quad A\neq\varnothing \tag{3-3-20}$$

是 mass 函数,其中

$$N=\sum_{E\cap F\neq\varnothing}m_1(E)m_2(F)>0 \tag{3-3-21}$$

为归一化系数。

证明 $m(\varnothing)=0$ 已经给定,只需证明 $\sum_{A\subseteq\mathcal{H}}m(A)=1$。而

$$\begin{aligned}\sum_{A\subseteq\mathcal{H}}m(A)&=m(\varnothing)+\sum_{A\subseteq\mathcal{H},A\neq\varnothing}m(A)=\sum_{A\subseteq\mathcal{H},A\neq\varnothing}\frac{1}{N}\sum_{E\cap F=A}m_1(E)\cdot m_2(F)\\&=\frac{1}{N}\sum_{E\cap F\neq\varnothing}m_1(E)\cdot m_2(F)=\frac{1}{N}N=1\end{aligned}$$

从而 m 是 mass 函数。 ■

设 m_1,m_2 是$\mathcal{H}$上的两个 mass 函数,而 m 是其合成的 mass 函数,一般记为

$$m=m_1\oplus m_2 \tag{3-3-22}$$

合成公式满足结合律和交换律。

一般情况下，如果$\mathcal{H}$上有 n 个 mass 函数 $m_1,m_2,\cdots,m_n$，如果

$$N=\sum_{\bigcap_{i=1}^{n}E_i\neq\varnothing}\prod_{i=1}^{n}m_i(E_i)>0 \tag{3-3-23}$$

则有如下合成公式：

$$m(A)=(m_1\oplus\cdots\oplus m_n)(A)=\frac{1}{N}\sum_{\bigcap_{i=1}^{n}E_i=A}\prod_{i=1}^{n}m_i(E_i) \tag{3-3-24}$$

如果 $N=0$，则所得 mass 函数存在矛盾。

例题 3.3.6 在例题 3.3.3 中，求两个医生诊断的 mass 函数，以及相应的信度函数和似真度函数。此时

$$\begin{aligned}N&=\sum_{E\cap F\neq\varnothing}m_1(E)\cdot m_2(F)=m_1(h)m_2(\mathcal{H})+m_1(\mathcal{H})m_2(\bar{h})+m_1(\mathcal{H})m_2(\mathcal{H})\\&=0.9\times0.8+0.1\times0.2+0.1\times0.8=0.82\end{aligned}$$

所以

$$m(h)=\frac{1}{0.82}m_1(h)m_2(\mathcal{H})=\frac{0.9\times0.8}{0.82}=\frac{36}{41}$$

$$m(\bar{h})=\frac{1}{0.82}m_1(\mathcal{H})m_2(\bar{h})=\frac{0.1\times0.2}{0.82}=\frac{1}{41}$$

$$m(\mathcal{H})=\frac{1}{0.82}m_1(\mathcal{H})m_2(\mathcal{H})=\frac{0.1\times0.8}{0.82}=\frac{4}{41}$$

这就是得到的 mass 函数。而按式(3-3-15)和式(3-3-16)，则 h 和 $\bar{h}$ 的信度函数与似真度函数分别为

$$Bel(h)=\sum_{D\subseteq h}m(D)=m(h)=\frac{36}{41}$$

$$Pl(h)=\sum_{D\cap h\neq\varnothing}m(D)=m(h)+m(\mathcal{H})=\frac{36+4}{41}=\frac{40}{41}$$

$$Bel(\bar{h})=\sum_{D\subseteq\bar{h}}m(D)=m(\bar{h})=\frac{1}{41}$$

$$Pl(\bar{h})=\sum_{D\cap\bar{h}\neq\varnothing}m(D)=m(\bar{h})+m(\mathcal{H})=\frac{1+4}{41}=\frac{5}{41}$$

所以 h 的信度区间为 $\left[\frac{36}{41},\frac{40}{41}\right]$，而 $\bar{h}$ 的信度区间为 $\left[\frac{1}{41},\frac{5}{41}\right]$。 □

例题 3.3.7 假定设备的故障有四种类型构成假设空间$\mathcal{H}=\{h_1,h_2,h_3,h_4\}$，而检测获取的系统状态估计分别是 $z_1,z_2\in\mathcal{O}$。现在已知给定 z_i 时的 mass 函数如下

$$m(\{h_1,h_2\}\mid z_1)=0.9;\quad m(\{h_3,h_4\}\mid z_1)=0.1$$

$$m(h_1\mid z_2)=0.7;\quad m(\{h_2,h_3,h_4\}\mid z_2)=0.3$$

（此时隐含：当 $A\neq\{h_1,h_2\}$或$\{h_3,h_4\}$时，$m(A\mid z_1)=0$；当 $A\neq\{h_1\}$或$\{h_2,h_3,h_4\}$时，$m(A\mid z_2)=0$）。

假定 z_1, z_2 发生的概率分别是 $\mu(z_1)=0.8, \mu(z_2)=0.2$，则

$$m(h_1)=m(h_1 \mid z_2)\mu(z_2)=0.7\times 0.2=0.14$$

$$m(\{h_1,h_2\})=m(\{h_1,h_2\} \mid z_1)\mu(z_1)=0.9\times 0.8=0.72$$

$$m(\{h_3,h_4\})=m(\{h_3,h_4\} \mid z_1)\mu(z_1)=0.1\times 0.8=0.08$$

$$m(\{h_2,h_3,h_4\})=m(\{h_2,h_3,h_4\} \mid z_2)\mu(z_2)=0.3\times 0.2=0.06$$

所以是 mass 函数。于是可得

$$Bel(h_1)=\sum_{D\subseteq h_1} m(D)=m(h_1)=0.14$$

$$Pl(h_1)=\sum_{D\cap h_1\neq\varnothing} m(D)=m(h_1)+m(\{h_1,h_2\})=0.14+0.72=0.86$$

$$Bel(\{h_1,h_2\})=\sum_{D\subseteq\{h_1,h_2\}} m(D)=m(h_1)+m(\{h_1,h_2\})=0.14+0.72=0.86$$

$$Pl(\{h_1,h_2\})=\sum_{D\cap\{h_1,h_2\}\neq\varnothing} m(D)=m(h_1)+m(\{h_1,h_2\})+m(\{h_1,h_2,h_3\})$$

$$=0.14+0.72+0.06=0.92$$

$$Bel(\{h_3,h_4\})=\sum_{D\subseteq\{h_3,h_4\}} m(D)=m(\{h_3,h_4\})=0.08$$

$$Pl(\{h_3,h_4\})=\sum_{D\cap\{h_3,h_4\}\neq\varnothing} m(D)=m(\{h_3,h_4\})+m(\{h_2,h_3,h_4\})$$

$$=0.08+0.06=0.14$$

$$Bel(\{h_2,h_3,h_4\})=\sum_{D\subseteq\{h_2,h_3,h_4\}} m(D)=m(\{h_2,h_3,h_4\})=0.06$$

$$Pl(\{h_2,h_3,h_4\})=\sum_{D\cap\{h_2,h_3,h_4\}\neq\varnothing} m(D)=m(\{h_1,h_2\})+m(\{h_3,h_4\})+m(\{h_2,h_3,h_4\})$$

$$=0.72+0.08+0.06=0.86$$

从而 h_1 的信度区间是$[0.14,0.86]$；$\{h_1,h_2\}$的信度区间是$[0.86,0.92]$；$\{h_3,h_4\}$的信度区间是$[0.08,0.14]$；而$\{h_2,h_3,h_4\}$的信度区间是$[0.14,0.86]$。 □

3.3.4 证据推理

定理 3.3.5 设 m 是假设空间$\mathcal{H}$上的 mass 函数，$\mathcal{P}(\mathcal{H})$表示$\mathcal{H}$的所有子集构成的幂集，P 是$\mathcal{H}$上的概率分布，则有 v：$\mathcal{P}(\mathcal{H})\to[0,1]$，满足 $v(\varnothing)=0$，且

$$v(A)=\frac{m(A)\cdot P(A)}{\sum\limits_{\varnothing\neq B\subseteq\mathcal{H}} m(B)\cdot P(B)} \tag{3-3-25}$$

以及 γ：$\mathcal{P}(\mathcal{H})\to[0,1]$，满足 $\gamma(\varnothing)=0$；且

$$\gamma(A)=\frac{m(A)/P(A)}{\sum\limits_{\varnothing\neq B\subseteq\mathcal{H}} m(B)/P(B)} \tag{3-3-26}$$

仍是$\mathcal{H}$上的 mass 函数。

证明 由于 $v(\varnothing)=0$，且

$$\sum_{\varnothing\neq A\subseteq\mathcal{H}} v(A)=\sum_{\varnothing\neq A\subseteq\mathcal{H}}\frac{m(A)\cdot P(A)}{\sum\limits_{\varnothing\neq B\subseteq\mathcal{H}} m(B)\cdot P(B)}=1$$

从而 v 是$\mathcal{H}$上的 mass 函数。同理可证 γ 也是$\mathcal{H}$上的 mass 函数。 ■

定理 3.3.6 设$\mathcal{H}$是有限集合，m_1 和 m_2 都是定义在$\mathcal{P}(\mathcal{H})$上的 mass 函数，P 是$\mathcal{H}$上的概率分布，则有 $v:\mathcal{P}(\mathcal{H})\to[0,1]$，满足 $v(\varnothing)=0$，且

$$v(A)=\frac{(\gamma_1\oplus\gamma_2)(A)\cdot P(A)}{\sum\limits_{\varnothing\neq D\subseteq\mathcal{H}}(\gamma_1\oplus\gamma_2)(D)\cdot P(D)},\quad A\in\mathcal{P}(\mathcal{H}) \tag{3-3-27}$$

仍是$\mathcal{H}$上的 mass 函数，其中

$$\gamma_i(A)=\frac{m_i(A)/P(A)}{\sum\limits_{\varnothing\neq B\subseteq\mathcal{H}} m_i(B)/P(B)},\quad i=1,2 \tag{3-3-28}$$

记为 $v(A)=(m_1\otimes m_2)(A)$。同时对于$\forall A\subseteq\mathcal{H}, A\neq\varnothing$，有

$$(m_1\otimes m_2)(A)=\frac{\sum\limits_{E\cap F=A}\left[P(A)\dfrac{m_1(E)}{P(E)}\cdot\dfrac{m_2(F)}{P(F)}\right]}{\sum\limits_{\varnothing\neq D\subseteq\mathcal{H}}\sum\limits_{E\cap F=D}\left[P(D)\dfrac{m_1(E)}{P(E)}\cdot\dfrac{m_2(F)}{P(F)}\right]} \tag{3-3-29}$$

证明 由定理 3.3.5 知，$m_1\otimes m_2$ 是$\mathcal{H}$上的 mass 函数；当$\forall A\subseteq\mathcal{H}, A\neq\varnothing$时，有

$$(m_1\otimes m_2)(A)=\frac{(\gamma_1\oplus\gamma_2)(A)\cdot P(A)}{\sum\limits_{\varnothing\neq B\subseteq\mathcal{H}}(\gamma_1\oplus\gamma_2)(B)\cdot P(B)}$$

代入 $\gamma_1\oplus\gamma_2$ 的合成公式并进行简化后得

$$(m_1\otimes m_2)(A)=\frac{\sum\limits_{E\cap F=A}[P(A)\cdot\gamma_1(E)\cdot\gamma_2(F)]}{\sum\limits_{\varnothing\neq D\subseteq\mathcal{H}}\sum\limits_{E\cap F=D}[P(D)\cdot\gamma_1(E)\cdot\gamma_2(F)]}$$

将式(3-3-28)代入上式做适当化简即得式(3-3-29)。 ■

归纳起来，证据推理的一般模型可用如下方式描述。

定义 3.3.6 设$\mathcal{H}=\{h_1,\cdots,h_n\}$表示假设空间，$\mathcal{P}(\mathcal{H})$表示$\mathcal{H}$的所有子集构成的集类；$\mathcal{O}=\{z_1,z_2,\cdots,z_l\}$表示观测空间，$\mu_i(z_i)$表示 z_i 发生的概率，而 P 是$\mathcal{H}$上的先验概率分布。对于每个 $z_i\in\mathcal{O}$，$m(\cdot\mid z_i)$和 $m(\cdot\mid\bar{z}_i)$都是$\mathcal{H}$上的 mass 函数，此处 $\bar{z}_i$ 表示 z_i 不发生。证据推理模型描述为

$$\{(\mathcal{H},P),(z_i,\mu_i),i=1,2,\cdots,l;\quad m(\cdot\mid z_i),m(\cdot\mid\bar{z}_i),i=1,2,\cdots,l\} \tag{3-3-30}$$

其中规则描述如下：如果观测 z_i 发生，则假设 $A\in\mathcal{P}(\mathcal{H})$具有强度 $m(A\mid z_i)$；如果观测 z_i 不发生，则假设 $A\in\mathcal{P}(\mathcal{H})$具有强度 $m(A\mid\bar{z}_i)$。

证据推理一般模型的计算步骤是：

(1) 计算 mass 函数 $m_i(A)$，即

$$m_i(A)=m(A\mid z_i)\mu_i(z_i)+m(A\mid\bar{z}_i)\mu_i(\bar{z}_i),\quad i=1,2,\cdots,l \tag{3-3-31}$$

(2) 利用先验概率 P，计算 mass 函数 $\gamma_i(A)$，即

$$\gamma_i(A)=\frac{m_i(A)/P(A)}{\sum\limits_{\varnothing\neq B\subseteq\mathcal{H}} m_i(B)/P(B)},\quad i=1,2,\cdots,l \tag{3-3-32}$$

(3) 利用 Dempster-Shafer 合成公式,计算 mass 函数 $\gamma(A)$,即

$$\gamma(A)=(\gamma_1 \oplus \gamma_2 \oplus \cdots \oplus \gamma_l)(A) \tag{3-3-33}$$

(4) 计算 mass 函数 $v(A)$,即

$$v(A)=\frac{\gamma(A)\cdot P(A)}{\sum\limits_{\varnothing \neq B \subseteq \mathcal{H}} \gamma(B)\cdot P(B)} \tag{3-3-34}$$

(5) 计算信度函数和似真度函数,即

$$Bel(A)=\sum_{B\subseteq A} v(B) \tag{3-3-35}$$

$$Pl(A)=\sum_{A\cap B\neq \varnothing} v(B) \tag{3-3-36}$$

于是,$[Bel(A),Pl(A)]$就是假设 $A\in \mathcal{P}(\mathcal{H})$的信任区间。

3.3.5 证据理论中的不确定度指标

描述不确定性的理论与方法很多,比如概率论、模糊集、可能性理论、证据理论以及粗糙集。这些理论可以看作是互补的,分别从不同角度来描述各种不同类型的不确定性。不确定性主要有三类[10],如图 3-3-1 所示。

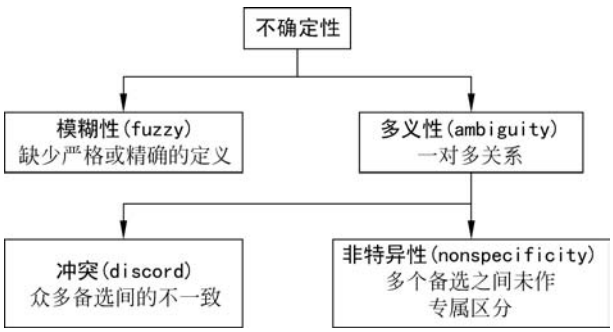


图 3-3-1 不确定性类型

本节将介绍几种用于描述不确定性程度的度量或指标——**不确定度**(**measure of uncertainty**)。

众所周知,严格的概率测度必须满足概率公理中的可列可加性。任何一个事件(或集合)的概率,必须等于这个事件的所有互斥子事件的概率之和。概率的这种特性就称为协调性。在经典集合论及概率框架中的不确定度主要有两类:

(1) Hartley 熵[11]:

$$\mathrm{H}(A)=\log_2(|A|),\quad A\subseteq \Theta \tag{3-3-37}$$

其中$|\cdot|$表示集合的势,即该集合包含元素的个数。

(2) Shannon 熵[12]:设 $p=\{p_\theta \mid u\in\Theta\}$是 Θ 上定义的概率分布,则熵定义为

$$\mathrm{S}(p)=-\sum_{\theta\in\Theta} p_\theta \log_2 p_\theta \tag{3-3-38}$$

这两个熵是建立其他不确定度的基础。对于证据理论而言,mass 函数不再遵守可

列可加性，即一个事件（或集合）的 mass 函数，不一定等于这个事件的所有互斥子事件的 mass 函数之和。对 mass 函数来说，可定义的不确定度主要包括：

（1）非特异度（nonspecificity measure）[13]：

$$\mathrm{N}(m)=\sum_{A\subseteq\Theta}m(A)\log_2|A| \tag{3-3-39}$$

非特异度是针对非特异性的度量，可看作是所有焦元的 Hartley 熵的加权和。当一个证据体的所有焦元均为单点时，mass 函数退化为概率，非特异性度量达到最小值 0。当证据体仅有一个焦元 Θ 时，非特异性度量达到最大值。

（2）聚合不确定性度（aggregated uncertainty measure，AU measure）[14]：设 Θ 为一有限论域，Bel 是定义在 Θ 上的信度函数。与 Bel 关联的聚合不确定度定义为

$$\mathrm{AU}(Bel)=\max_{\mathcal{P}_{Bel}}\left[-\sum_{\theta\in\Theta}p_\theta\log_2 p_\theta\right] \tag{3-3-40}$$

这实际上是一个优化求解问题，即在所有可能的概率分布情况下，选取使得式（3-3-40）取值最大的一组概率分布；这组概率分布要求与给定的信度函数 Bel 一致，即满足 $p_\theta\in[0,1]$，$\forall\theta\in\Theta$ 且 $\sum_{\theta\in\Theta}p_\theta=1$；同时要求 $Bel(A)=\sum_{\theta\in\Theta}p_\theta$，$\forall A\subseteq\Theta$。式（3-3-40）中的 $\mathcal{P}_{Bel}$ 代表所有满足上述条件的概率分布。

AU 满足作为不确定性度的五个条件[15]：

① 概率一致性条件：当 m 定义为概率测度，则 AU 蜕化为 Shannon 信息熵。

② 集合一致性条件：当证据体的焦元只有一个的时候，即 $\exists A\subseteq\Theta$，$m(A)=1$，$\forall B\neq A$，$m(B)=0$，AU 蜕化为 Hartley 熵，即 $\mathrm{AU}(m)=\log_2|A|$。

③ 取值范围：$0\leqslant\mathrm{AU}(m)\leqslant\log_2|\Theta|$。

④ 边缘分布的次可加性条件：即对任意两个基本概率赋值 m_X 与 m_Y，其联合基本概率赋值 m 满足：$\mathrm{AU}(m)\leqslant\mathrm{AU}(m_X)+\mathrm{AU}(m_Y)$。

⑤ 独立边缘分布的可加性条件：即对任意两个独立的基本概率赋值 m_X 与 m_Y，其联合基本概率赋值 m 满足：$\mathrm{AU}(m)=\mathrm{AU}(m_X)+\mathrm{AU}(m_Y)$。

AU 的设计也有些不足之处[8]，如计算复杂度、对证据变化的高度不敏感性等。

（3）总体不确定度（total uncertainty measure，TU measure）：可以看作是非特异度和 AU 的线性组合，即

$$\mathrm{TU}(m,\delta)=\delta\mathrm{AU}(m)+(1-\delta)\mathrm{N}(m) \tag{3-3-41}$$

3.3.6 证据理论存在的主要问题与发展

证据理论在表示和处理不确定性问题时具有明显的优势，然而，在处理不确定性推理时也存在一些问题[17-18]，主要集中在 Dempster 证据组合公式的使用上。

（1）组合爆炸问题。Dempster 证据组合规则最为直观的一个缺陷在于在进行证据组合时，会引起“焦元爆炸”问题，焦元数目随着辨识框架元素数目以指数形式递增，造成计算量的激增。如果辨识框架有 n 个元素，那么可能的焦元数目为 2^n+n-1 个。以 20 个元素的辨识框架为例，所有可能焦元的数目为 1.048576×10^7。针对这一问题，已经出现了不少解决方法，主要的思路包括预设焦元结构、减少焦元数量以及改进证据组合算

法等。比较有代表性的是基于神经网络[19]的证据组合算法以及 Barnett 提出的方法[20]。

(2) 有限辨识框架及证据体独立性问题。Dempster 证据组合规则虽然具有简单的表达式,但其隐含条件是相对苛刻的,实际应用中往往忽略或者弱化这些条件,这样的后果是会带来很多不合理现象。证据组合公式的使用条件包括:

① D-S 证据理论讨论的是一种离散且有限的辨识框架(close-world),这种框架是由一些具有完备性和互斥性(exhaustive and exclusive)的元素构成;

② 参与证据组合的证据体,被认为是相互独立的。

正是这样两个条件的存在,限制了 Dempster 组合规则在一些不确定性问题中的应用。针对条件①中说到的辨识框架问题,Schubert 在 *On nonspecific evidence*[21]一文中做过讨论。下面介绍的 DSmT 可以说是对传统证据理论框架的一个拓展,能够从很大程度上解决因辨识框架引起的问题。条件②中提及的证据体独立性问题,在实际应用中往往很难满足,即绝对严格独立的证据体几乎是不存在的。很多研究者致力于研究对相关证据的处理[22-24],对相关证据的 mass 函数进行去相关性处理,然后再进行证据组合。有学者提出了证据能量、相关系数等概念[25],并对组合规则做了改进,也得到了一些有意义的结论。

(3) 高冲突证据组合问题。当参与证据组合的证据体之间存在较大冲突时,Dempster 组合规则往往无法有效处理,常常会得出有悖常理、违反直觉的结论。关于这方面的研究,已经成了当前热点之一。下面将针对这一问题做出详细的讨论。需要指出的是,这里讨论的"冲突"(conflict)是指参与证据组合的证据体之间的冲突,而上节所述的"冲突"(discord)是针对证据体内部而言的。

此外,在证据理论的实际使用当中,mass 函数的生成或获取往往是最困难的一步,也有很多文献致力于设计合理有效的 mass 函数生成法。Dempster 证据组合规则是对称的规则,参与证据组合的证据体没有重要性上的差异。许多研究工作都围绕建立非对称的证据推理规则而展开,特别是条件证据与证据更新。

基于 Dempster-Shafer 证据理论,许多研究者都作了拓展性研究,提出了一些新的模型和规则,诸如 TBM 模型[26]、DSmT[27]等。

(1) 可传递信度模型。可传递信度模型(transferable belief model,TBM)是 Philippe Smets 于 20 世纪 80 年代,在证据理论基础上提出的一种扩展模型。TBM 旨在建立一种用不完全概率函数来量化信度表示的理论,主要研究信度随证据变化时的更新问题。

TBM 模型是一个双层结构,包括 credal 层以及 pignistic 层。其中,在 credal 层处理的是信度的获取和更新问题;在 pignistic 层,将 credal 层得到的信度转换成 pignistic 概率进而据此进行决策。credal 层先于 pignistic 层。在 credal 层上随时可以对信度进行赋值和更新,在 pignistic 层仅进行决策。

① credal 层:在 credal 层对于证据处理的规则包括 Dempster 证据组合规则以及式(3-3-42)中的条件推理规则。

定义 3.3.7[26] 假设 m 是辨识框架 Θ 上的 mass 函数。若新到的证据支持 Θ 上的一个命题 B,即 $m(B)=1$,则定义条件 mass 函数为

$$m_B(A)=\begin{cases}\dfrac{\sum\limits_{X\subseteq\bar{B}}m(A\cup X)}{1-\sum\limits_{X\subseteq\bar{B}}m(X)}, & A\subseteq B\\ 0, & \text{其他}\end{cases} \tag{3-3-42}$$

从上式中不难看出：当新出现的证据支持命题（或事件）B 为真，则原有证据中的基本信度赋值 $m(A)$ 传递到命题 $A\cap B$ 上。可传递信度模型也正是因此而得名。

② pignistic 层：在 credal 层上做证据处理时，焦元可以是单点的（singleton），也可以是复合的（compound or composite）。所谓单点焦元对应于辨识框架中的一个单一元素（单命题）。而复合焦元对应于辨识框架中元素的组合（复合命题）。依据单点焦元给出的结论是清晰的，而依据复合焦元给出的结论相对是模糊的。证据组合实质上就是一种聚焦过程，使得模糊的结论能够相对清晰化。在 TBM 模型中，决策时需要将 credal 层上的各种单命题及复合命题结论都转换成单命题结论，也就是将基于证据理论的信度框架转换到概率框架来进行决策。为此，Smets 定义了 pignistic 概率：

$$\mathrm{BetP}_m(A)=\sum_{B\subseteq\Theta}m(B)\frac{|A\cap B|}{|B|} \tag{3-3-43}$$

对于单点焦元，是

$$\mathrm{BetP}_m(\theta)=\sum_{\theta\in B,B\subseteq\Theta}\frac{m(B)}{|B|} \tag{3-3-44}$$

其中，$|\cdot|$ 表示集合中元素的个数。从式（3-3-43）可以看出，一个复合焦元的 mass 函数赋值是被平均分配到组成该复合焦元的所有单点焦元上的。

（2）Dezert-Smarandache 理论。由 Dezert 和 Smarandache 等创立的 Dezert-Smarandache 理论（DSmT）[27-29]，又称作**似真与冲突推理理论（plausible and paradoxical reasoning）**，其主要创新是在辨识框架中引入了冲突信息。相对于 D-S 证据理论中的幂集 2^Θ 概念，DSmT 提出了**超幂集（hyper-powerset）**D^Θ 的概念。D^Θ 是通过辨识框架 $\Theta=\{\theta_1,\theta_2,\cdots,\theta_n\}$ 中的基本元素进行交、并运算产生的集合；即在 DSmT 中，辨识框架中的元素不再是不可分的。D^Θ 需要满足以下三个条件：

① $\varnothing,\theta_1,\theta_2,\cdots,\theta_n\in D^\Theta$；

② 若 $A,B\in D^\Theta$，则 $A\cap B\in D^\Theta$ 且 $A\cup B\in D^\Theta$；

③ 除了条件①、②中获得的元素之外，再没有其他元素属于 D^Θ。

对于 DSmT 来说，由于它是建立在超幂集的概念上的，运算量是个令人头疼的问题，当 $|\Theta|=n$ 时，D^Θ 的势达到 2^{2^Θ} 数量级。

DSmT 是从 D-S 证据理论基础上拓展而来，因此也沿袭使用了 D-S 证据理论中的基本概率赋值（mass 函数）、信度函数（Bel）以及似真度函数（Pl）等。但由于模型基本框架的变化，DSmT 的组合规则不同于 D-S 证据理论中的 Dempster 组合规则。

定义 3.3.8　（DSmT 证据组合规则）[27]　假定辨识框架 Θ 上性质不同的两个证据，其焦元分别为 B_i 和 C_j（$i=1,2,\cdots,n$；$j=1,2,\cdots,m$），其 mass 函数分别为 m_1,m_2，组合后的 mass 函数为

$$m(A)=\begin{cases}0, & A=\varnothing\\ \sum\limits_{B_i,C_j\in D^{\Theta},B_i\cap C_j=A} m_1(B_i)m_2(C_j), & A\neq\varnothing\end{cases} \tag{3-3-45}$$

这个规则在实际中有广泛应用。

近年来证据理论又有很大发展。证据距离用来度量两条证据之间的差异程度，证据距离越大，则两条证据差异程度越大。下面给出一些新的概念。

(1) Tessem 距离[88]。

$$d_{\mathrm{T}}(m_1,m_2)=\max_{A\subseteq\Theta}\{|\ \mathrm{Bet}P_1(A)-\mathrm{Bet}P_2(A)\ |\} \tag{3-3-46}$$

其中，Bet$P(A)$就是前述 pignistic 概率转换得到的转换概率[89]，定义如式(3-3-43)所示。

(2) 基于模糊隶属度的证据相似性度量[90]。

$$d_{\mathrm{F}}(m_1,m_2)=1-\frac{\sum\limits_{i=1}^{n}(\mu^{(1)}(\theta_i)\ \wedge\ \mu^{(2)}(\theta_i))}{\sum\limits_{i=1}^{n}(\mu^{(1)}(\theta_i)\ \vee\ \mu^{(2)}(\theta_i))} \tag{3-3-47}$$

其中，$\wedge$ 表示合取运算(最小值)，$\vee$ 表示析取运算(最大值)。

(3) Jousselme 距离[91]。

$$d_{\mathrm{J}}(m_1,m_2)=\sqrt{0.5\cdot(m_1-m_2)^{\mathrm{T}}Jac(m_1-m_2)} \tag{3-3-48}$$

其中，$Jac(A,B)$是 Jaccard 加权矩阵，$Jac(A,B)=\dfrac{|A\cap B|}{|A\cup B|}$。

(4) 基于区间数的证据距离[92]。

假设区间数表示为 $\bar{a}$，用 a^- 表示 $\bar{a}$ 的下界，a^+ 表示 $\bar{a}$ 的上界。区间数 $\bar{a}_1$ 与 $\bar{a}_2$ 之间的 Wasserstein 距离[6]定义为

$$d_{BI}(\bar{a}_1,\bar{a}_2)=\sqrt{\left[\frac{a_1^-+a_1^+}{2}-\frac{a_2^-+a_2^+}{2}\right]^2+\frac{1}{3}\left[\frac{a_1^+-a_1^-}{2}-\frac{a_2^+-a_2^-}{2}\right]^2} \tag{3-3-49}$$

两个 mass 函数 $m_i(i=1,2)$的信度区间模型如下：

$$BI_i=\begin{bmatrix}[Bel_i(A_1),Pl_i(A_1)]\\ [Bel_i(A_2),Pl_i(A_2)]\\ \vdots\\ [Bel_i(A_{2^n-1}),Pl_i(A_{2^n-1})]\end{bmatrix}=\begin{bmatrix}BI_i(A_1)\\ BI_i(A_2)\\ \vdots\\ BI_i(A_{2^n-1})\end{bmatrix}$$

基于 Wasserstein 距离定义了两种证据距离，Euclidean 系列的证据距离定义为

$$d_{BI}^{E}(m_1,m_2)=\sqrt{N_{\mathrm{c}}\cdot\sum_{i=1}^{2^n-1}[d_{BI}(BI_1(A_i),BI_2(A_i))]^2} \tag{3-3-50}$$

其中，$N_{\mathrm{c}}=1/2^{n-1}$ 是归一化因子。Chebyshev 系列的证据距离定义为

$$d_{BI}^{C}(m_1,m_2)=\max_{A_i\subseteq\Theta,i=1,\cdots,2^n-1}\{d_{BI}(BI_1(A_i),BI_2(A_i))\} \tag{3-3-51}$$

3.3.7 关于证据函数不确定性的研讨

关于证据的不确定性是一个值得进一步研讨的问题，如下资料供研究者参考。证据函数的不确定性度量包含两个部分，分别是不一致性（discord）和非特异性（non-specificity）。常用的不确定性度量方法中有分别针对不一致性和非特异性进行度量的，也有对这两部分进行总体度量的。

1. 关于不一致性的度量方法

（1）Confusion 度量（confusion measure）[96]

$$\mathrm{Conf}(m)=-\sum_{A\subseteq\Theta}m(A)\log_2(Bel(A)) \tag{3-3-52}$$

（2）Dissonance 度量（dissonance measure）[97]

$$\mathrm{Diss}(m)=-\sum_{A\subseteq\Theta}m(A)\log_2(Pl(A)) \tag{3-3-53}$$

（3）Discord 度量（discord measure）[98]

$$\mathrm{Disc}(m)=-\sum_{A\subseteq\Theta}m(A)\log_2\left[1-\sum_{B\subseteq\Theta}m(B)\frac{|B-A|}{|B|}\right] \tag{3-3-54}$$

（4）Strife 度量（strife measure）[99]

$$\mathrm{Stri}(m)=-\sum_{A\subseteq\Theta}m(A)\log_2\left[1-\sum_{B\subseteq\Theta}m(B)\frac{|A-B|}{|A|}\right] \tag{3-3-55}$$

2. 关于非特异性的度量方法

（1）Dubois & Prade 非特异性度量[100]

$$\mathrm{NS}^{\mathrm{DP}}(m)=\sum_{A\subseteq\Theta}m(A)\log_2|A| \tag{3-3-56}$$

（2）Yager 特异性度量[97]

$$S^{\mathrm{Y}}(m)=\sum_{A\subseteq\Theta}\frac{m(A)}{|A|} \tag{3-3-57}$$

（3）Korner 非特异性度量[101]

$$\mathrm{NS}^{\mathrm{K}}(m)=\sum_{A\subseteq\Theta}m(A)\cdot|A| \tag{3-3-58}$$

（4）Korner 非特异性度量（一般形式）[101]

$$\mathrm{NS}^{\mathrm{f}}(m)=\sum_{A\subseteq\Theta}m(A)\cdot f(|A|) \tag{3-3-59}$$

3. 总体不确定性度量方法

（1）聚合不确定度（aggregated uncertainty measure，AU）[102]

$$\mathrm{AU}(m)=\max\left[-\sum_{\theta=\Theta}p_\theta\log_2 p_\theta\right] \tag{3-3-60}$$

$$\text{s.t.}\begin{cases}\sum\limits_{\theta\in\Theta} p_\theta = 1\\ p_\theta \in [0,1], \quad \forall \theta \in \Theta\\ Bel(A) \leqslant \sum\limits_{\theta\in\Theta} p_\theta \leqslant 1 - Bel(\overline{A}), \quad \forall A \in \Theta\end{cases}$$

(2) 含混度(ambiguity measure, AM)[103]

$$\mathrm{AM}(m) = -\sum_{\theta\in\Theta} \mathrm{Bet}P_m(\theta)\log_2(\mathrm{Bet}P_m(\theta)) \tag{3-3-61}$$

其中, $\mathrm{Bet}P_m(\theta) = \sum\limits_{\theta\in B\subseteq\Theta} m(B)/|B|$, 是 pignistic 概率转换得到的转换概率[105]。

(3) 基于距离的总体不确定性度量[106]

$$\mathrm{TU}^I(m) = 1 - \frac{1}{n}\cdot\sqrt{3}\cdot\sum_{i=1}^{n} d_{\mathrm{BI}}([Bel(\{\theta_i\}), Pl(\{\theta_i\})], [0,1]) \tag{3-3-62}$$

其中, d^I 是信度区间 $[Bel(\{\theta_i\}), Pl(\{\theta_i\})]$ 和 $[0,1]$ 之间的 Wasserstein 距离[44]。$[0,1]$是最不确定的情况。Wasserstein 距离能够表示两区间数之间的距离。区间数 $\tilde{a}$, 用 a^- 表示 $\tilde{a}$ 的下界, a^+ 表示 $\tilde{a}$ 的上界。区间数 $\tilde{a}_1$ 与 $\tilde{a}_2$ 之间的 Wasserstein 距离定义为

$$d_{\mathrm{BI}}(\tilde{a}_1, \tilde{a}_2) = \sqrt{\left[\frac{a_1^- + a_1^+}{2} - \frac{a_2^- + a_2^+}{2}\right]^2 + \frac{1}{3}\left[\frac{a_1^+ - a_1^-}{2} - \frac{a_2^+ - a_2^-}{2}\right]^2} \tag{3-3-63}$$

3.4 随机集理论基础

自从 G. Matheron 于 1975 年出版了《随机集与积分几何学》一书以来,关于随机集理论的研究方兴未艾[31-34],其成功地用于解决各种问题,现在已经成为信息融合研究领域最受关注的研究方向之一。

3.4.1 一般概念

从本质上讲,随机集与随机变量并没有太大的差别,随机变量处理的是随机点问题,而随机集处理的是随机集合问题,即后者处理的对象是可能结果的一组元素。如果把后者的集值结果看作是某个特殊空间的点,则二者就没有区别了。然而,集值的引入却带来许多令人意想不到的奇特结果,这正是利用随机集建立多源信息融合统一理论方面最吸引人的地方。

考虑一个概率空间$(\Omega, \mathcal{F}, P)$和一个可测空间$(Y, \mathcal{B}_Y)$,其中 Y 是观测集合,而$\mathcal{B}_Y$ 表示相应的 Borelσ-域。对于一般的随机变量 $x: \Omega \to Y$ 而言,概率空间只是一个数学上的概念,而实际观测到的变量在可测空间。定义在概率空间上的概率测度 P 一般难以直接用于计算,因此通常把它转换成可测空间上的概率分布,即

$$P_x \triangleq P \circ x^{-1}: \mathcal{B}_Y \to [0,1] \tag{3-4-1}$$

也就是说,要求 x 可测,即对于 $\forall A \in \mathcal{B}_Y$,则 $\{\omega \in \Omega: x(\omega) \in A\} \in \mathcal{F}$,而且

$$P_x(A) = P\{\omega: x(\omega) \in A\} \tag{3-4-2}$$

下面考虑一个集值映射。

定义 3.4.1 设$(\Omega,\mathcal{F},P)$是一个概率空间，$(Y,\mathcal{B}_Y)$是一个可测空间，定义**集值映射**(**set-valued mapping**)为

$$X:\Omega\to\mathcal{P}(Y) \tag{3-4-3}$$

其中$\mathcal{P}(Y)$是Y的所有子集构成的集类。给定$A\in\mathcal{B}_Y$，其**上逆**(**upper inverse**)定义为

$$X^*(A)\triangleq\{\omega\in\Omega:X(\omega)\cap A\neq\varnothing\} \tag{3-4-4}$$

下逆(**lower inverse**)定义为

$$X_*(A)\triangleq\{\omega\in\Omega:\varnothing\neq X(\omega)\subseteq A\} \tag{3-4-5}$$

逆(**inverse**)定义为

$$X^{-1}(A)\triangleq\{\omega\in\Omega:X(\omega)=A\} \tag{3-4-6}$$

上述集值映射称为是**开的**(或**闭的**或**紧的**)，如果对于任意$\omega\in\Omega$，$X(\omega)$是Y中**开的**(或**闭的**或**紧的**)子集。

在不致引起混淆的前提下，我们记

$$A^*\triangleq X^*(A),\quad A_*\triangleq X_*(A),\quad A^{-1}\triangleq X^{-1}(A) \tag{3-4-7}$$

一个集值映射的上逆和下逆，实质上是对随机变量逆概念的推广，显然有$A_*\subseteq A^*$。

定理 3.4.1 设$X:\Omega\to\mathcal{P}(Y)$是一个集值映射，则如下条件等价：

① $X(\omega)\neq\varnothing,\quad\forall\omega\in\Omega$ (3-4-8)

② $X_*(A)\subseteq X^*(A),\quad\forall A\subseteq Y$ (3-4-9)

③ $X_*(\varnothing)=\varnothing$ (3-4-10)

④ $X^*(Y)=\Omega$ (3-4-11)

证明 见文献[30]。 ■

定理 3.4.2 设$X:\Omega\to\mathcal{P}(\Omega)$是一个集值映射，则如下条件等价：

① $\omega\in X(\omega),\quad\forall\omega\in\Omega$ (3-4-12)

② $X_*(A)\subseteq A,\quad\forall A\subseteq\Omega$ (3-4-13)

③ $A\subseteq X^*(A),\quad\forall A\subseteq\Omega$ (3-4-14)

证明 见文献[30]。 ■

定义 3.4.2 设$X:\Omega\to\mathcal{P}(Y)$是一个集值映射，定义算子：$j:\mathcal{P}(Y)\to\mathcal{P}(\Omega)$，且对于任意$A\in\mathcal{P}(Y)$，记

$$j(A)=X^{-1}(A)\subseteq\Omega \tag{3-4-15}$$

如果$j(A)\neq\varnothing$，则称A是j的一个**焦集**(**focus set**)；且令

$$\mathcal{J}\triangleq\{j(A)\in\mathcal{P}(\Omega):j(A)\neq\varnothing,A\subseteq Y\} \tag{3-4-16}$$

则必有

$$\begin{cases}\bigcup\{j(A):A\subseteq Y\}=\Omega\\ A_1,A_2\in\mathcal{P}(Y),A_1\neq A_2\Rightarrow j(A_1)\cap j(A_2)=\varnothing\end{cases} \tag{3-4-17}$$

这样，$\mathcal{J}$构成对Ω的一个划分，于是称算子j是集值映射X的**关系划分函数**。

定理 3.4.3 设$\mathcal{P}_0(Y)=\mathcal{P}(Y)-\varnothing$，$X:\Omega\to\mathcal{P}_0(Y)$是一个集值映射，$j$是$X$的关系划分函数，则有如下性质：

① $X_*(A)=\bigcup\{j(A'):A'\subseteq A\},\quad A\subseteq Y$ (3-4-18)

② $X^*(A)=\bigcup\{j(A'):A'\cap A\neq\varnothing\},\quad A\subseteq Y$ (3-4-19)

③ $j(A)=X_*(A)-\bigcup\{X_*(A'):A'\subseteq A\},\quad A\subseteq Y$ (3-4-20)

证明 见文献[30]。 ■

从概念上说,一个随机集就是满足某些可测性条件的集值映射。虽然根据不同的可测性有不同的定义,但大多数都是基于上逆和下逆的概念。

定义 3.4.3 式(3-4-3)定义的集值映射 X 如果对于 $\forall A\in\mathcal{B}_Y$ 均有 $A^*\in\mathcal{F}$,则称 X 是**强可测的(strongly measurable)**;而强可测的集值映射 X 定义为**随机集(random set)**。

注意,对于 $\forall A\in\mathcal{B}_Y$,因为 $A_*=Y^*\cap((A^c)^*)^c$(注 A^c 是 A 在 Y 中的补集,$(B^*)^c$ 是 B^* 在 Ω 中的补集),所以如果 X 是强可测的,必然对于 $\forall A\in\mathcal{B}_Y$ 均有 $A_*\in\mathcal{F}$。

随机集是对随机变量概念的推广,是由基本事件空间到观测空间的幂集的一个映射,它比随机变量高了一个复杂层次。随机集不再符合概率公理的一切结论。然而,在解决复杂系统问题时就可能克服所遇到的困难,这种对随机变量的概念进行推广,在一个复杂层次上可以满足系统分析的需要。

定义 3.4.4 给定一个随机集 $X:\Omega\to\mathcal{P}(Y)$,$A\in\mathcal{B}_Y$ 的**上概率(upper probability)**定义为

$$P_X^*(A)\triangleq P(A^*)/P(Y^*) \tag{3-4-21}$$

下概率(lower probability)定义为

$$P_{*X}(A)\triangleq P(A_*)/P(Y^*) \tag{3-4-22}$$

在不致因为不确定哪个随机集诱导上概率和下概率而引起混淆的前提下,记

$$P^*=P_X^*,\quad P_*=P_{*X} \tag{3-4-23}$$

我们可以把随机集视为对一个随机变量 $x_0:\Omega\to Y$ 不精确观察的结果,这个随机变量就称为**原始随机变量(original random variable)**,而且对于任意 $\omega\in\Omega$,$x_0(\omega)\in X(\omega)$。因此,假定对于任意 $\omega\in\Omega$,$X(\omega)\neq\varnothing$,从而对 $\forall A\in\mathcal{B}_Y$ 有 $P^*(A)=P(A^*)$,$P_*(A)=P(A_*)$。这样,由随机集诱导的上概率和下概率是一对共轭函数。

如果随机集 X 是对随机变量 x_0 不精确观察的结果,我们有关这个随机变量的所有知识就是它属于 X 的**可测选择类(class of measurable selection)**或称为**选择器(selector)**,即

$$S_X\triangleq\{x:\Omega\to Y\text{ 是随机变量},x(\omega)\in X(\omega),\forall\omega\} \tag{3-4-24}$$

x_0 的概率分布属于

$$P_X\triangleq\{P_x:x\in S_X\} \tag{3-4-25}$$

关于 $P_{x_0}(A)$ 的信息由如下**值集(set of values)**给出

$$P_X(A)\triangleq\{P_x(A):x\in S_X\},\quad\forall A\in\mathcal{B}_Y \tag{3-4-26}$$

还有另外两个概率类在某些场合是有用的,其中第一个是

$$\Delta_X\triangleq\{Q\text{ 是概率分布}:Q(A)\in P_X(A),\forall A\in\mathcal{B}_Y\} \tag{3-4-27}$$

这是概率分布函数的集合,其值是与随机集给出的信息相容的,显然有 $P_X\subseteq\Delta_X$。如果它们相一致,关于原始变量分布的信息等价于其取值的信息;而第二个概率类是

$$M_{P^*} \triangleq \{Q \text{ 是概率分布：} Q(A) \leqslant P^*(A), \forall A \in \mathcal{B}_Y\} \tag{3-4-28}$$

这是由 P^* 支配的概率分布函数，或者说是由 P^* 生成的**纲领集**（**credal set**）。这个概率类是凸的，在实际应用中比 P_X 容易处理。因为不等式 $P_*(A) \leqslant P_x(A) \leqslant P^*(A)$，对于任意 $x \in S_X$，$A \in \mathcal{B}_Y$ 都有效，我们能够推导出 $\Delta_X \subseteq M_{P^*}$，然后得到 $P_X \subseteq \Delta_X \subseteq M_{P^*}$。可以证明，这两个包含关系可能是严格包含的，所以利用上概率和下概率在某些场合会使精度降低，而反过来也会引起某种错误判定。因此，我们感兴趣的是判断在什么情况下利用 P_* 和 P^* 是合理的。

虽然一个随机集的选择器的分布类和由它诱导的上概率在许多文献中都深入研究过，但必须注意它们之间的联系。对于 Y 是有限集合的情况，下面将特别予以关注。我们将着重讨论以下两个问题：

（1）研究 Δ_X 与 M_{P^*} 之间的关系使我们知道，对于某些任意的 $A \in \mathcal{B}_Y$，上概率和下概率关于 $P_{x_0}(A)$ 的值是否提供足够的信息。

（2）研究什么时候 $P_X = M_{P^*}$，即在什么条件下上概率保持关于 $P_{x_0}(A)$ 的所有信息。

3.4.2 概率模型

1. $P_*(A)$，$P^*(A)$ 作为 $P_{x_0}(A)$ 的模型

首先，我们研究 Δ_X 与 M_{P^*} 之间的关系。因为我们已经提到，Δ_X 是对某些信息的建模，而这些信息给出了 $\mathcal{B}_Y$ 中元素的概率值。因此，通过研究其与 M_{P^*} 的相等关系，则可窥见对于任意 $A \in \mathcal{B}_Y$ 的“真”概率，P_* 和 P^* 的信息是充足的。

定理 3.4.4 设 $(\Omega, \mathcal{F}, P)$ 是一个概率空间，$(Y, \mathcal{B}_Y)$ 是一个可测空间，$X: \Omega \to \mathcal{P}(Y)$ 是随机集，则

$$\Delta_X = M_{P^*} \Leftrightarrow P_X(A) = [P_*(A), P^*(A)], \quad \forall A \in \mathcal{B}_Y \tag{3-4-29}$$

证明 见文献[35]。 ■

任取 $A \in \mathcal{B}_Y$，然后考虑 $P_X(A)$ 和 $[P_*(A), P^*(A)]$，显然后者是前者的扩展集合。为了给出相等条件，我们必须考察 $P_X(A)$ 的最大、最小值是否和 $P^*(A)$，$P_*(A)$ 相一致，而且 $P_X(A)$ 是否是凸的。

文献中已经证明，$P_X(A)$ 有一个最大值和一个最小值（实际上是 $[0,1]$ 区间的一个闭子集），而且这些值并不在任意情况下与 $P^*(A)$，$P_*(A)$ 相一致，即使在 $S_X \neq \varnothing$ 的非平凡情况下也是如此。进而，在一般情况下 $P_X(A)$ 也不是凸的。下面的定理给出 $P^*(A) = \max P_X(A)$ 和 $P_*(A) = \min P_X(A)$ 的充分条件。

定理 3.4.5 考虑 $(\Omega, \mathcal{F}, P)$ 是一个概率空间，(Y, d) 是一个度量空间，$X: \Omega \to \mathcal{P}(Y)$ 是随机集，则在满足下列任意条件的前提下：

① Y 是可分的度量空间，且 X 是紧的；

② Y 是可分的度量空间，且 X 是开的；

③ Y 是光滑的空间，且 X 是闭的；

④ Y 是 σ-紧度量空间，且 X 是闭的；

则有

$$P^*(A)=\max P_X(A),\quad P_*(A)=\min P_X(A),\quad \forall A\in\mathcal{B}_Y \tag{3-4-30}$$

证明 （略）。■

注 该定理对于等式 $P^*=\max P_X$，以及 $P_*=\min P_X$ 给出了充分条件。P^* 的一致性意味着它是所支配有限相加概率集合的上包络。可以证明，在定理规定条件下，它实际上就是选择器诱导的可数相加概率类的上包络。类似地，可以对 P_* 做出类似的评论。

定理 3.4.6 设 $X:\Omega\to\mathcal{P}(Y)$ 是一个随机集，$v:Y\to\mathbb{R}$ 是一个有界随机变量，只要满足定理 3.4.5 的几个条件之一，则

$$\int v\mathrm{d}P^*=\sup\left\{\int v\mathrm{d}P_x:x\in S_X\right\},\quad \int v\mathrm{d}P_*=\inf\left\{\int v\mathrm{d}P_x:x\in S_X\right\} \tag{3-4-31}$$

证明 （略）。■

定理 3.4.7 设 $X:\Omega\to\mathcal{P}(Y)$ 是一个随机集，且 $A\in\mathcal{B}_Y$，令 $x_1,x_2\in S_X$ 满足

$$P_{x_1}(A)=\max P_X(A),\quad P_{x_2}(A)=\min P_X(A)$$

则

$$P_X(A)\text{ 是凸的}\Leftrightarrow x_1^{-1}(A)-x_2^{-1}(A)\text{ 不是不可分的最小元} \tag{3-4-32}$$

证明 （略）。■

注 此处所谓不可分的最小元是指对于 $\forall\alpha\in(0,1)$，不存在任何可测集 $B\in\mathcal{F}$，使得 $B\in[x_1^{-1}(A)-x_2^{-1}(A)]$，且 $P(B)=\alpha P[x_1^{-1}(A)-x_2^{-1}(A)]$。

推论 设 $X:\Omega\to\mathcal{P}(Y)$ 是一个随机集，满足定理 3.4.5 的几个条件之一。如果对于 $\forall A\in\mathcal{B}_Y$，$A^*-A_*$ 不是不可分的最小元，则

$$\Delta_X=M_{P^*} \tag{3-4-33}$$

2. P^*,P_* 作为 P_{x_0} 的模型

现在研究 P_X 与 M_{P^*} 之间的相等关系，因为这使我们知道上概率是否保持了原始随机变量分布 P_{x_0} 的所有可用的信息。Δ_X 与 M_{P^*} 之间的相等关系，并不能保证 P_X 与 M_{P^*} 之间存在相等关系。那么一种可行的方法就是确定 $P_X=\Delta_X$ 的充分条件，然后结合上述推论得到所要的结果。不幸的是推论中的式子判定却并不容易。在有限维情况下，表征 P_X 与 M_{P^*} 之间的相等关系将更容易，可由此导出 Y 是可分度量空间条件下的一些有用结果。

当 Y 是有限集时，给定 $Y=\{y_1,y_2,\cdots,y_n\}$，$X:\Omega\to\mathcal{P}(Y)$ 是一个随机集，可以得出：$P_X=M_{P^*}\Leftrightarrow P_X$ 是凸的。但是这个等价关系对于 Y 由无限个元构成的集合并不成立。

例题 3.4.1 设 $Y=\{y_1,y_2,\cdots,y_n\}$ 是一个有限集，$X:\Omega\to\mathcal{P}(Y)$ 是一个随机集，而 $\mathcal{P}(Y)$ 由 2^n 个 Y 的可能子集构成（包括空集 $\varnothing$ 和 Y 本身），则 $X(\omega)\in\mathcal{P}(Y)$，$\forall\omega\in\Omega$。设 S^n 表示序列 $\{1,2,\cdots,n\}$ 所有可能的重排构成的集合，而 $\pi\in S^n$，则有序列 $\{\pi(1),\pi(2),\cdots,\pi(n)\}$ 是对原序列的一种重排。对于任意 $A\in\mathcal{P}(Y)$，则 $\exists\pi\in S^n$，$j=0,1,2,\cdots,n$，使得

$$A=\{y_{\pi(1)},y_{\pi(2)},\cdots,y_{\pi(j)}\};\quad \text{当 } j=0 \text{ 时，}\quad A=\varnothing \tag{3-4-34}$$

于是有

$$\begin{cases} P_X(A)=P_X(\{y_{\pi(1)},y_{\pi(2)},\cdots,y_{\pi(j)}\}) \\ P^*(A)=\max P_X(A) \\ P_*(A)=\min P_X(A) \end{cases} \tag{3-4-35}$$

现在考虑 $n=3$ 的情况，则

$Y=\{y_1,y_2,y_3\}$，　$\mathcal{P}(Y)=\{\varnothing,\{y_1\},\{y_2\},\{y_3\},\{y_1,y_2\},\{y_1,y_3\},\{y_2,y_3\},Y\}$

对于 $\forall A\in\mathcal{P}(Y)$，$P_x(A)$=事件发生的概率。不妨记为

$$\begin{cases} P_x(\varnothing)=p_{000}, & P_x(\{y_1\})=p_{100} \\ P_x(\{y_2\})=p_{010}, & P_x(\{y_3\})=p_{001} \\ P_x(\{y_1,y_2\})=p_{110}, & P_x(\{y_1,y_3\})=p_{101} \\ P_x(\{y_2,y_3\})=p_{011}, & P_x(Y)=p_{111} \end{cases} \tag{3-4-36}$$

则有表 3-4-1 所示的上概率和下概率。

表 3-4-1　$Y=\{y_1,y_2,y_3\}$ 情况下的上概率和下概率

$A\in\mathcal{P}(Y)$	$P^*(A)$	$P_*(A)$
$\varnothing$	0	0
$\{y_1\}$	$(p_{100}+p_{110}+p_{101}+p_{111})/(1-p_{000})$	$p_{100}/(1-p_{000})$
$\{y_2\}$	$(p_{010}+p_{110}+p_{011}+p_{111})/(1-p_{000})$	$p_{010}/(1-p_{000})$
$\{y_3\}$	$(p_{001}+p_{101}+p_{011}+p_{111})/(1-p_{000})$	$p_{001}/(1-p_{000})$
$\{y_1,y_2\}$	$(p_{100}+p_{010}+p_{110}+p_{101}+p_{011}+p_{111})/(1-p_{000})$	$(p_{100}+p_{010}+p_{110})/(1-p_{000})$
$\{y_1,y_3\}$	$(p_{100}+p_{001}+p_{110}+p_{101}+p_{011}+p_{111})/(1-p_{000})$	$(p_{100}+p_{001}+p_{101})/(1-p_{000})$
$\{y_2,y_3\}$	$(p_{010}+p_{001}+p_{110}+p_{101}+p_{011}+p_{111})/(1-p_{000})$	$(p_{010}+p_{001}+p_{011})/(1-p_{000})$
Y	1	1

此时，$P_X(A)=[P_*(A),P^*(A)]\subseteq[0,1]$。　□

例题 3.4.2　设 $X:(0,1)\to\mathcal{P}([0,1])$ 是一个随机集，把概率空间 $((0,1),\mathcal{B}_{(0,1)},\lambda_{(0,1)})$ 映射到可测空间 $([0,1],\mathcal{B}_{[0,1]})$，且 $X(\omega)=(0,\omega)$，　$\forall\omega\in(0,1)$。容易看出这个映射是强可测的。

(1) 给定 $x\in S(X)$，可以验证 $P_x(\{0\})=0$，$P_x([0,\omega])\geqslant\omega$，$\forall\omega\in(0,1)$，而且 $\lambda_{(0,1)}(\{\omega\in(0,1):P_x([0,\omega])=\omega\})=0$。

(2) 反之，考虑一个概率测度 $Q:\mathcal{B}_{[0,1]}\to[0,1]$ 满足 $Q([0,\omega))\geqslant\omega$，而且对于 $(0,1)$ 的几乎空子集(all but a null subset) N_Q，有 $Q([0,\omega))>\omega$。Q 的分位点函数(quantile function) x 是一个可测映射，满足 $P_x=Q$，$x(\omega)\in X(\omega)$，$\forall\omega\notin N_Q$。对 N_Q 上的 x 进行修改，对其可测性不影响，使得它的取值包含在 X 的值中，据此能够导出 $Q\in P_X$。

(3) 能够推导出 P_X 是概率测度类，且 $Q(\{0\})=0$，$Q([0,\omega])\geqslant\omega$，$\forall\omega$；而且对于 $[0,1]$ 的几乎空子集，有 $Q([0,\omega])>\omega$，而且容易验证这个类是凸的。

(4) $\mathcal{B}_{[0,1]}$ 上的 Lebesgue 测度 $\lambda_{[0,1]}$ 满足 $\lambda_{[0,1]}(A)\leqslant P^*(A)$，$\forall A\in\mathcal{B}_{[0,1]}$，因此它属于 M_{P^*}，而且显然并不以概率 1 满足 $\lambda_{[0,1]}([0,\omega])>\omega$，因而 $P_X\not\subset M_{P^*}$。　□

定理 3.4.8　设 (Y,d) 是一个可分的度量空间，考虑一个集类 $\mathcal{U}\subseteq\mathcal{B}_Y$，使得

(1) 对有限交运算是封闭的;

(2) 每个开集都是有限集,或者是$\mathcal{U}$中元的可数并,令$\{P_n\}$和P都是$\mathcal{B}_Y$上的概率测度,使得

$$P_n(A) \underset{n\to\infty}{\to} P(A), \quad \forall A \in \mathcal{U} \tag{3-4-37}$$

那么,序列$\{P_n\}$弱收敛于P。

证明 (略)。 ■

3.4.3 随机集的 mass 函数模型

本节将在随机集理论框架内对 mass 函数模型重新描述,以便得到二者之间的联系。

定理 3.4.9 考虑$(\Omega,\mathcal{F},P)$是一个概率空间,U是一个有限集合构成的论域空间,而$X:\Omega\to\mathcal{P}(U)$是随机集,且对$\forall\omega\in\Omega, X(\omega)\neq\varnothing, X(\omega)\neq U$,则

$$m(A)=P\{X^{-1}(A)\} \tag{3-4-38}$$

$$Bel(A)=P\{X_*(A)\} \tag{3-4-39}$$

$$Pl(A)=P\{X^*(A)\} \tag{3-4-40}$$

分别是$\mathcal{P}(U)$上的 mass 函数、信度函数与似真度函数,而且

$$Be(A)=\sum_{A'\subseteq A} m(A') \tag{3-4-41}$$

$$Pl(A)=\sum_{A'\cap A\neq\varnothing} m(A') \tag{3-4-42}$$

于是对任意$A\in\mathcal{P}(U)$,$Bel(A)\leqslant Pl(A)$,$Bel(A)=1-Pl(A^c)$。

证明 因为$X(\omega)\neq\varnothing$,则$X^{-1}(\varnothing)=\varnothing$;同时当$A_1\neq A_2$时,$X^{-1}(A_1)\cap X^{-1}(A_2)=\varnothing$;且有

$$\sum_{A\subseteq Y} X^{-1}(A)=U$$

于是$X^{-1}(A)$是对Ω的关系划分,从而m是 mass 函数。又因为

$$Bel(A)=\sum_{A'\subseteq A} m(A')=\sum_{A'\subseteq A} P(X^{-1}(A'))=P\Big(\sum_{A'\subseteq A} X^{-1}(A')\Big)$$

$$=P(\omega\in\Omega:\varnothing\neq X(\omega)\subseteq A)=P(X_*(A))$$

所以式(3-4-39)得证;类似可证式(3-4-40)。由信度函数与似真度函数的定义即可得到式(3-4-41)和式(3-4-42)。 ■

这样,我们把研究随机集的问题与研究证据理论的问题就统一起来了。

定理 3.4.10 考虑$(\Omega,\mathcal{F},P)$是一个概率空间,U是一个有限集合构成的论域空间,而$X_1:\Omega\to\mathcal{P}(U)$,$X_2:\Omega\to\mathcal{P}(U)$是相互独立的随机集,相应的 mass 函数分别是$m_1$和$m_2$,则 Dempster-Shafer 合成公式就是这两个相互独立随机集的交运算,即

$$m(A)=P\{\omega\in\Omega: X_1(\omega)\cap X_2(\omega)=A\}, \quad A\in\mathcal{P}(U) \tag{3-4-43}$$

证明 设m_1,m_2是$\mathcal{P}(U)$上的两个独立的 mass 函数,即存在两个独立的随机集$X_1:\Omega\to\mathcal{P}(U)$和$X_2:\Omega\to\mathcal{P}(U)$,使得

$$m_1(E)=P\{X_1^{-1}(E)\}, \quad m_2(F)=P\{X_2^{-1}(F)\}, \quad \forall E,F\subseteq U$$

任意取 m_1，m_2，假定 $A=\varnothing$，则 $m(\varnothing)=0$ 已经给定；当 $A\neq\varnothing$，选择相互独立的 $E,F\subseteq U$，使得 $E\cap F=A$，则两个 mass 函数的合成就意味着

$$\begin{aligned}m(A)&=P(X^{-1}(A))=P\{\omega\in\Omega:X(\omega)=A\}\\&=\frac{1}{N}\sum_{E\cap F=A}P\{\omega\in\Omega:X_1(\omega)=E,X_2(\omega)=F\}\\&=\frac{1}{N}\sum_{E\cap F=A}P\{\omega\in\Omega:X_1(\omega)=E\}\cdot P\{\omega\in\Omega:X_2(\omega)=F\}\\&=\frac{1}{N}\sum_{E\cap F=A}P(X_1^{-1}(E))\cdot P(X_2^{-1}(F))=\frac{1}{N}\sum_{E\cap F=A}m_1(E)\cdot m_2(F)\end{aligned}$$

其中 N 是归一化因子。现在只需证明，即

$$\begin{aligned}\sum_{B\subseteq U}m(B)&=m(\varnothing)+\sum_{B\subseteq U,B\neq\varnothing}m(B)=\sum_{B\subseteq U,B\neq\varnothing}\frac{1}{N}\sum_{E\cap F=B}m_1(E)\cdot m_2(F)\\&=\frac{1}{N}\sum_{E\cap F\neq\varnothing}m_1(E)\cdot m_2(F)=\frac{1}{N}N=1\end{aligned}$$

从而 m 是 mass 函数。 ■

注 应用如此广泛且理论推导如此晦涩的证据合成理论，用随机集的观点来看竟然是如此简单，仅仅是两个独立随机集的交运算。

Dempster 公式的意义在于，两个不同的可能性判断经过合成后变成统一的判断，这是一种形式的信息融合，在许多实际问题中都有应用。注意，这是一种集合运算，与概率的数值计算完全不同。一般情况下，如果 U 上有 n 个独立的 mass 函数 $m_1,m_2,\cdots,m_n$，而且 $N=\sum_{\bigcap_{i=1}^{n}E_i\neq\varnothing}\prod_{i=1}^{n}m_i(E_i)>0$，则有如下合成公式

$$m(A)=(m_1\oplus\cdots\oplus m_n)(A)=\frac{1}{N}\sum_{\bigcap_{i=1}^{n}E_i=A}\prod_{i=1}^{n}m_i(E_i)\tag{3-4-44}$$

如果 $n=0$，则所得 mass 函数存在矛盾。

对于非独立的几个 mass 函数的合成，见文献[36]。

3.4.4 随机集与模糊集的转换

证据理论是不确定性表示和推理的重要理论和方法之一。对于不确定性推理的定量方法，首先要解决的问题是如何实现对不确定性信息的表示和度量。不同的不确定性推理方法各有其不同的不确定性信息表示与描述方法。证据理论中的不确定性信息表示与描述可通过 mass 函数来实现。定义在某一辨识框架幂集上的 mass 函数是对传统概率定义的一种推广。确定一个 mass 函数应包括两个部分：焦元结构的确定和相应的 mass 赋值的确定。得到 mass 函数，即获得随机集描述，便可以很方便地求取证据理论中的其他函数（信度函数、似真度函数）以及依据证据组合规则进行不确定性推理。在证据理论的实际应用中，mass 函数的生成与获取至关重要，却往往也是最困难的一步。目

前还没有统一的方法，其生成方法往往与具体应用相关。本节将讨论 mass 函数生成方法。

定义 3.4.5 设$(\Omega,\mathcal{F},P)$是一个概率空间，U 是一个有限集合构成的论域空间，而 $X:\Omega\to\mathcal{P}(U)$是随机集，定义

$$\mu_X(u)\triangleq P\{\omega\in\Omega:u\in X(\omega)\},\quad \forall u\in U \tag{3-4-45}$$

称之为该随机集的**单点覆盖函数(one-point covering function)**。

单点覆盖函数 $\mu_X:U\to[0,1]$可以视为一个模糊隶属函数，这样就把模糊集 μ_X 与随机集 X 联系起来了。注意，这是由随机集(或 mass 函数)向模糊集的转换，而反过来由模糊集向随机集(或 mass 函数)的转换要复杂得多。

用模糊集诱导随机集合的方法可以很多，先介绍一种简单的方法。

定义 3.4.6 设 $\widetilde{F}$ 是 U 上的一个模糊集，ω 是区间$[0,1]$上均匀分布的随机变量，则相应的随机集可以定义为

$$X_{\widetilde{F}}(\omega)\triangleq\widetilde{F}^{-1}[\omega,1] \tag{3-4-46}$$

其中 $\widetilde{F}^{-1}:A\subseteq[0,1]\to\mathcal{P}(U)$表示逆映射，从而 $X_{\widetilde{F}}:[0,1]\to\mathcal{P}(U)$是一个随机集。

这样，研究模糊集的问题就变成研究随机集的问题。但是，这种转换赋予一个很强的假设：ω 是区间$[0,1]$上均匀分布。这个假设在很多情况下不一定成立。

我们研究 mass 函数生成的一般方法。

1. 依据目标类型数和环境加权系数确定 mass 函数[37-38]

当考虑目标类型数及传感器(信息源)环境对分类和识别的影响时，可采用如下的经验方法确定 mass 函数。

设 M 为传感器目标类型数，N 是传感器总数，$C_i(o_j)$是传感器 i 对目标类型 o_j 的关联系数($j=1,\cdots,M,i=1,\cdots,N$)。$\lambda_i$ 是传感器 i 的环境加权系数，且定义

$$\begin{cases}\alpha_i=\max\{C_i(o_j)\mid j=1,\cdots,M\}\\ \xi_i=M\lambda_i\Big/\sum_{j=1}^{M}C_i(o_j)\\ \beta_i=(\xi_i-1)/(N-1),N\geqslant 2\\ R_i=(\lambda_i\alpha_i\beta_i)\Big/\sum_{l=1}^{N}\lambda_l\alpha_l\beta_l\end{cases},\quad i=1,\cdots,N \tag{3-4-47}$$

传感器(信息源)i 对目标 o_j 的 mass 函数转换为

$$\begin{cases}m_i(\{o_j\})=\dfrac{C_i(o_j)}{\sum_{l=1}^{M}C_i(o_l)+M(1-R_i)(1-\lambda_i\alpha_i\beta_i)}\\ m_i(\{\Theta\})=\dfrac{M(1-R_i)(1-\lambda_i\alpha_i\beta_i)}{\sum_{l=1}^{M}C_i(o_l)+M(1-R_i)(1-\lambda_i\alpha_i\beta_i)}\end{cases},\quad i=1,\cdots,N \tag{3-4-48}$$

上式是对辨识框架全集的基本概率赋值，用来描述“不知道”或“未知”。

在本方法中，焦元结构预先设定为单点焦元和辨识框架的全集。

2. 基于统计证据的 mass 函数获取方法[37,39]

若有一批证据是基于统计实验结果获得的，称这批证据为**统计证据**。统计证据是证据理论在统计问题中的应用，是 Shafer 用非统计方法研究统计问题的一种尝试。

设统计试验的观测由一组概率模型$\{p_\theta|\theta\in\Theta\}$确定。$p_\theta$ 是给定 θ 时的概率密度函数。Shafer 的方法给出下面两个假设。

第一个假设：观测 $\boldsymbol{x}$ 决定一个似真度函数，满足

$$Pl(\{\theta\})=cp_\theta(\boldsymbol{x}),\quad \forall\theta\in\Theta \tag{3-4-49}$$

其中，c 为常数，与 θ 无关。

第二个假设：似真度函数应该是一致(consonant)的。所谓“一致”是指所有证据体中焦元都是嵌套(nested)结构的，即对于任一 $A_i\subseteq\Theta, i=1,\cdots,r$，总有：$m(A_i)>0$ 及 $\sum_{i=1,\cdots,r} m(A_i)=1$，$A_i\subset A_j,\forall i<j$。

在上述两个假设成立的前提下，得到的一致似真度函数为

$$Pl(A)=\frac{\max\{p_\theta:\theta\in A\}}{\max\{p_\theta:\theta\in\Theta\}},\quad A\neq\varnothing, A\subseteq\Theta \tag{3-4-50}$$

结合式(3-4-49)可得

$$c=\frac{1}{\max\{p_\theta:\theta\in\Theta\}} \tag{3-4-51}$$

相应的支撑函数为

$$Sp_x(A)=1-\frac{\max\{p_\theta:\theta\in A^c\}}{\max\{p_\theta:\theta\in\Theta\}} \tag{3-4-52}$$

假定 $\Theta^0=\{\theta^{(1)},\theta^{(2)},\cdots,\theta^{(N)}\}$ 是 Θ 的有序集，满足 $p_{\theta^{(i)}}>p_{\theta^{(j)}},\forall 1\leqslant i<j\leqslant N$，则称 $Sp_x(A)$ 为一致支撑函数。相应的 mass 函数为

$$m_x(A)=\begin{cases}\dfrac{p_{\theta^{(k)}}(\boldsymbol{x})-p_{\theta^{(k+1)}}(\boldsymbol{x})}{p_{\theta^{(1)}}(\boldsymbol{x})}, & \forall A=\{\theta^{(1)},\theta^{(2)},\cdots,\theta^{(k)}\},1\leqslant k\leqslant N-1\\ \dfrac{p_{\theta^{(N)}}(\boldsymbol{x})}{p_{\theta^{(1)}}(\boldsymbol{x})}, & \forall A=\Theta^0=\Theta\\ 0, & \text{其他}\end{cases} \tag{3-4-53}$$

上述一致 mass 函数虽然易于实现，但只适用于满足一致支持假设的特定场合。为此，可通过弱化一致支持假设得到推广方法：设$\{W_1,W_2,\cdots,W_r\}$是 Θ 的一个划分，若 $\forall k=1,\cdots,r$，Sp 在每个 W_k 上是一致的，我们说支持函数 $Sp:2^\Theta\to[0,1]$是部分一致的。若 Sp 在 Θ 的一个划分$\{W_1,W_2,\cdots,W_r\}$上部分一致，令 $W_k^o=\{w_k^{(1)},w_k^{(2)},\cdots,w_k^{(n_k)}\}$是 W_k 对应的有序集，使得 $\forall 1\leqslant i<j\leqslant n_k$，有 $p_{w_k^{(i)}}>p_{w_k^{(j)}}$。其中 $\sum_{k=1,\cdots,r} n_k=N$。此时，$Sp$ 对应的 mass 函数为

$$m(A)=\begin{cases}C_p\{p_{w_k^{(l)}}-p_{w_k^{(l+1)}}\}, & \forall A=\{w_k^{(1)},w_k^{(2)},\cdots,w_k^{(l)}\},1\leqslant l\leqslant n_k-1\\ C_p p_{w_k^{(n_k)}}, & \forall A=W_k^o,1\leqslant k\leqslant n_k-1\\ 0, & \text{其他}\end{cases} \tag{3-4-54}$$

其中，$C_p=\left[\sum_{k=1,\cdots,r}\max\{p_w:w\in W_k\}\right]^{-1}$。

3. 基于隶属度函数生成 mass 函数的方法

经典集合论中，“属于”或“不属于”，两者必取其一，非此即彼。也就是说经典集合中的特征函数与二值逻辑相对应，只能取 0，1 两个值，也就无法表达现实中存在的“亦此亦彼”的模糊现象。取值推广到[0，1]闭区间，这就形成了模糊数学中的隶属度函数。关于隶属度函数的获取已经出现了许多行之有效的方法，包括模糊统计方法、主观指派方法、二元对比排序法等[40-41]。

设 $\Theta=\{\theta_1,\theta_2,\cdots,\theta_n\}$ 为辨识框架，$\{f_1,f_2,\cdots,f_n\}$ 为相应的隶属度函数，有如下几种典型的 mass 函数生成方法。

(1) 依 $P_i=f_i\Big/\sum_{j=1,\cdots,n}f_j$ 将隶属度函数归一化，对应的 mass 函数生成如下[42]

$$m(\{\theta_i\})=P_i \tag{3-4-55}$$

该方法实际上是预设了焦元结构为单点焦元。这样的 mass 函数定义方式实际上等同于标准的概率定义。单点焦元的优势在于运算简单方便，但无法充分体现证据理论的优势，即无法表达和区分“不确定”和“不知道”。

(2) 依 $P_i=f_i\Big/\sum_{j=1,\cdots,n}f_j$ 将隶属度函数归一化，对应的 mass 函数生成如下[43]

$$\begin{cases}m(\{\theta_i\})=\alpha P_i\\ m(\Theta)=1-\alpha\sum_{i=1}^{n}P_i\end{cases} \tag{3-4-56}$$

其中，α 为折扣系数。该方法实际上是预设了焦元结构为单点焦元及辨识框架的全集。将概率值打折扣赋予单点焦元，剩余部分赋给全集，以表征“不知道”。

(3) 依 $P_i=f_i\Big/\sum_{j=1,\cdots,n}f_j$ 将隶属度函数归一化，对应的 mass 函数生成如下[43]

$$\begin{cases}m(\{\theta_i\})=\alpha\cdot P_i\\ m(\{\theta_i^c\})=\alpha\cdot\sum_{j\neq i}P_j\\ m(\Theta)=1-m(\{\theta_i\})-m(\{\theta_i^c\})=1-\alpha\end{cases} \tag{3-4-57}$$

其中，α 为折扣系数，$\theta_i^c=\Theta-\theta_i$ 是 θ_i 的余集。该方法实际上是将 mass 赋值赋予辨识框架中单点焦元及其相应反命题的焦元，剩余部分 mass 赋值赋予全集，来描述“不知道”。

(4) 依 $P_i=f_i\Big/\sum_{j=1,\cdots,n}f_j$ 将隶属度函数归一化，取 $i_1=\arg\max_j P_j$，$i_2=\arg$

$\max\limits_{j,j\neq i}P_j$，对应的 mass 函数生成如下[44]：

$$\begin{cases} m(A_1)=P_{i_1}, & A_1=\{\theta_{i_1}\} \\ m(A_2)=P_{i_2}, & A_2=\{\theta_{i_2}\} \\ m(\Theta)=1-m(A_1)-m(A_2) \end{cases} \tag{3-4-58}$$

该方法中，选取隶属度值最高的辨识框架中的两个元素，作为单点焦元，以及辨识框架全集作为焦元结构的定义。归一化过后的隶属度赋值的最大两个赋值赋予上述两个单点焦元，剩余赋值赋予全集，来描述“不知道”。

(5) 协调证据中的隶属度函数向 mass 函数的转换。协调的 mass 函数，也称作一致支撑的 mass 函数，焦元满足依次嵌套。此时隶属度函数等价于单点似真度函数[45-46]。此时的单点似真度函数也被称为外形函数或单点覆盖函数。

一般地，设 m 是一个协调（一致支持）的 mass 函数，若其焦元为

$$A_i=\{x_j;\ j\leqslant i\}\quad (i\leqslant n) \tag{3-4-59}$$

其中，n 为辨识框架中元素的个数。$r_i, i=1,\cdots,n, \max\limits_i r_i=1$ 为其隶属度函数，则有

$$m(A_i)=r_i-r_{i+1}\quad (i\leqslant n) \tag{3-4-60}$$

其中，$r_{n+1}=0$。

该方法求解 mass 函数，也预设了焦元结构为嵌套形式。

本节介绍的很多方法都是由隶属度函数出发来获取 mass 函数，这是因为隶属度函数本身是针对单点赋值的，其获取与 mass 函数的获取（包括焦元结构的确定以及相应的 mass 赋值的确定）相比，相对较为容易。

还有其他一些 mass 函数的获取方法，比如文献[47]中，利用目标速度和加速度获取 mass 函数；文献[48]中，利用模糊 c-均值聚类来实现 mass 函数的获取；而在文献[49]中，则是利用基于 Gauss 模型假设的迭代估计来实现 mass 函数获取；在文献[50]中，利用广义粗集模型中的容差关系及决策表来生成 mass 函数。随着证据理论研究的深入和应用的拓展，相信会有越来越多行之有效的 mass 函数获取方法出现。

3.5 随机有限集概略

因为随机有限集理论正在成为多目标跟踪领域的一个研究热点，而且其理论基础与随机集理论有一定的联系，因而本节主要讨论随机有限集的基本概念和理论，然后讨论 Mahler 的有限集统计理论。

3.5.1 概述

在图像研究中，如果物体几何形状是随机变化的，用一个点过程描述这样的问题就很难满足。在目标跟踪中，如果跟踪的目标不是一个点目标，而是一个随机变化的目标群，同样用点过程来描述和处理有本质的困难。

20 世纪 60 年代，人们已经注意到这个问题。Moyal 首先提出了群体随机点过程的概念[51]。1975 年，Mathéron 第一个系统地提出了随机几何的思想，并提出用随机集（闭

集)来解决这个问题的思路[52]。随机有限集是具有有限个元的随机集,由于其简单性在处理实际问题时受到特别重视,人们对其进行了更多的研究。1976 年 Ripley[53] 的研究和 1984 年 Baudin [54]的研究指出,随机有限集在本质上也是随机点过程的研究基础,二者实质上是等价的。随机有限集应用于信息融合领域起始于 1990 年 Mahler 提出的有限集统计(Finite Set Statistics,FISST)理论[55-61],并且逐步应用于目标跟踪领域[62]。

3.5.2 随机有限集的概念

先考虑 Mathèron 拓扑,Mathèron 拓扑又称为 Hit-or-Miss 拓扑。在实数域上,经常用到 Borel σ-代数。同样,在实数域上所有的闭子集,我们希望找到某种意义下的拓扑$\mathcal{T}$。在该意义下,随机闭集的概率属性能够用更为简单的方式刻画出来。

设 $U=\mathbb{R}^d$ 表示 d 维实数空间,$\mathcal{F}(\mathbb{R}^d)$表示$\mathbb{R}^d$ 上的所有闭子集,Ω 表示基本事件空间。定义随机集 $X:\Omega\to\mathcal{F}$,设 $A\subseteq U$,子集类$\mathcal{F}_A=\{F\in\mathcal{F}:F\cap A\neq\varnothing\}\subseteq\mathcal{F}$,指的是"击中"(Hit)$A$ 的所有闭子集类;而$\mathcal{F}_A$的补集类$\mathcal{F}_A^c=\{F\in\mathcal{F}:F\cap A=\varnothing\}$则表示"未击中"(Miss)$A$ 的所有子闭子集类。所以以这种方式产生的拓扑称为 Hit-or-Miss 拓扑。如果仅限于有限可测闭子集,那么这种映射就称为随机有限集,下面给出严格定义。

定义 3.5.1 对于局部紧的 Hausdorff 可分空间 E(例如$\mathbb{R}^n$),$\mathcal{F}(E)$是 E 上所有有限子集构成的集类,称为 Mathéron 拓扑;Ω 表示基本事件空间,定义可测映射[56,63]

$$\Sigma:\Omega\to\mathcal{F}(E) \tag{3-5-1}$$

那么,Σ 就称为**随机有限集(random finite sets,RFS)**。

考虑$(\Omega,\mathcal{B},P)$是一个概率空间,定义于概率测度 P 上的 RFS 的**概率特性**为

$$P_\Sigma(A)=P(\Sigma^{-1}(A))=P(\{\omega:\Sigma(\omega)\in A\}) \tag{3-5-2}$$

分别有三种方式进行刻画:

(1) **概率分布函数**。对于任意的 Borel 子集 $A\in\mathcal{F}(E)$,随机有限集 Σ 的分布函数为

$$P_\Sigma(A)=P(\Sigma^{-1}(A))=P(\{\omega:\Sigma(\omega)=A\})^{①} \tag{3-5-3}$$

(2) **信度函数(belief mass function,BMF)**。对于任意闭子集 $A\subseteq E$,

$$\beta_\Sigma(A)=P(\{\omega:\Sigma(\omega)\subseteq A\}) \tag{3-5-4}$$

(3) **空概率(void probability)**函数。对于任意闭子集 $A\subseteq E$,

$$\zeta_\Sigma(A)=P(\{\omega:\Sigma(\omega)\cap A\neq\varnothing\})=\beta_\Sigma(A^c) \tag{3-5-5}$$

其中,A^c 是A 的补集。

3.5.3 随机有限集的统计

对于一般的随机变量的概率密度(离散情况下是概率质量函数),可以通过积分(离散情况下求和)获得概率分布函数。同样,对概率分布函数可利用 Randon-Nikedym 导数获得概率密度函数。

但对于 RFS(随机有限集)来说,定义在可测集 E 上的置信测度不满足可列可加性,

① 参考文献[55]给出的表达式是 $P_\Sigma(A)=P(\Sigma^{-1}(A))=P(\{\omega:\Sigma(\omega)\in A\})$,疑是笔误。

Randon-Nikedym 导数不能直接应用，并且传统的概率分布是定义在区间上的，而对于集合来说，区间不再适用，只能用集合之间的包含关系来定义，因此随机有限集 Σ 的概率分布定义如下：

$$P_{\Sigma}(S)=P(\Sigma \in S) \tag{3-5-6}$$

随机有限集的测度和一般概率密度不同，它是定义在可测的有限集上的，而不是 Broel 集上。它的概率属性一般通过置信函数（belife-mass function）来刻画。我们以随机有限集的积分为例来说明随机有限集不能直接套用通常的概率统计理论。设 $Y\in\mathbb{R}^m$，它的期望定义为

$$E(Y)=\int_{S\subseteq\mathbb{R}^m}\boldsymbol{y}p_Y(y)\mathrm{d}\boldsymbol{y} \tag{3-5-7}$$

其中 $S\subseteq\mathbb{R}^m$ 是概率密度函数 $p_Y(\boldsymbol{y})$的分布范围。但是，对于随机有限集 Σ 来说，必须定义 $\mathbb{R}^m$ 上的某个 σ 有限测度 q，令 $Y=q(\Sigma)$，根据 Robbin 公式，Y 的期望定义为

$$E(Y)=E[q(\Sigma)]=\int\mu_{\Sigma}(\boldsymbol{z})\mathrm{d}q(\boldsymbol{z}) \tag{3-5-8}$$

其中 $\mu_{\Sigma}(\boldsymbol{z})$是随机有限集 Σ 的单点覆盖函数，并且 $\mu_{\Sigma}(z)\triangleq P(\omega\in\Omega:\boldsymbol{z}\in\Sigma(\omega))$。如果 Σ 为随机有限集，并且 $q(\boldsymbol{z})$是关于 Lebesgue 测度绝对连续的，就会发现对任何随机有限集，$q(\boldsymbol{z})=0$，并且 $\mu_{\Sigma}(\boldsymbol{z})=0$；因此，Robbin 公式只有平凡解。这说明对随机有限集来说，不能直接套用一般随机变量积分的思路。

由于类似的原因，以及随机有限集问题本身的复杂性，它在实际工程中很难直接应用。为此，Mahler 提出了一套面向工程的基于随机有限集的统计方法，即有限集统计理论（FISST）。其关键是解决随机有限集的置信密度和置信函数之间的关系，核心是建立了 FISST 上的有限集积分和有限集导数。

1. 有限集积分

定义 3.5.2 对于 RFS，$Y\subseteq\mathcal{Y}$，$\mathcal{Y}$是目标跟踪空间；对于 Y 的任意实值函数 $f(Y)$，在空间 $S\subseteq\mathcal{Y}$上的积分定义为

$$\int_S f(Y)\delta Y=f(\varnothing)+\sum_{n=1}^{\infty}\frac{1}{n!}\int_{S^n}f(\{\boldsymbol{y}_1,\cdots,\boldsymbol{y}_n\})\mathrm{d}\boldsymbol{y}_1\cdots\mathrm{d}\boldsymbol{y}_n \tag{3-5-9}$$

如果 $S=\mathcal{Y}$时满足$\int_{S=\mathcal{Y}}f(Y)\delta Y=1$，那么 $f(Y)$就是密度函数。

此时，似然函数和转移密度同样满足$\int f_k(Z\mid X)\delta Z=1$，$\int f_{k+1|k}(Y\mid X)\delta Y=1$。

例题 3.5.1（全局均匀密度函数）[64] 考虑混杂空间$\mathcal{R}=\mathbb{R}^n\times W$，其中 W 是有限的离散空间。T 为混杂空间$\mathcal{R}$上的一个闭子集，对于所有的有限子集 $Z\subseteq\mathcal{R}$，定义如下集函数为

$$u_{T,M}(Z)=\begin{cases}\dfrac{|Z|!}{M\lambda(T)^{|Z|}}, & Z\subseteq T,|Z|<M\\ 0, & \text{其他}\end{cases}$$

那么根据式(3-5-9)，对于所有闭子集 S 有

$$\begin{aligned}\int_S u_{T,M}(Z)\delta Z &= \sum_{n=0}^{\infty}\frac{1}{n!}\int_{S^n} u_{T,M}(\{\xi_1,\cdots,\xi_n\})\mathrm{d}\lambda(\xi_1)\cdots\mathrm{d}\lambda(\xi_n)\\ &= \sum_{n=0}^{M-1}\frac{1}{n!}\int_{(T\cap S)^n}\frac{n!}{M\lambda(T)^n}\mathrm{d}\lambda(\xi_1)\cdots\mathrm{d}\lambda(\xi_n)\\ &= \frac{1}{M}\sum_{n=0}^{M-1}\frac{\lambda(S\cap T)^n}{\lambda(T)^n}=1\end{aligned}$$ □

2. 有限集导数

对于任意随机有限集 Y 的集函数 $F(Y)$，其导数定义为

$$\begin{cases}\dfrac{\delta F}{\delta \boldsymbol{y}}(S)=\lim\limits_{\nu(E_y)\to 0}\dfrac{F(S\cup E_y)-F(S)}{\nu(E_y)}\\ \dfrac{\delta\beta}{\delta Y}(S)=\dfrac{\delta^n\beta}{\delta\boldsymbol{y}_n\cdots\delta\boldsymbol{y}_1}(S)=\dfrac{\delta\delta^{n-1}\beta}{\delta\boldsymbol{y}_n\delta\boldsymbol{y}_{n-1}\cdots\delta\boldsymbol{y}_1}(S)\end{cases}\tag{3-5-10}$$

其中 $\beta(S)$ 是置信函数，$\nu(S)$ 是状态空间 S 的超体积(Lebesgue 测度)。

有限集导数满足如下的和律、积律、常数律和幂律。

和律：
$$\frac{\delta}{\delta Y}(a_1\beta_1(S)+a_2\beta_2(S))=a_1\frac{\delta\beta_1}{\delta Y}(S)+a_2\frac{\delta\beta_2}{\delta Y}(S)\tag{3-5-11}$$

其中 a_1,a_2 是任意常数。

积律：
$$\frac{\delta}{Y}(\beta_1(S)\beta_2(S))=\sum_{W\subseteq Y}\frac{\delta\beta_1}{\delta W}(S)\frac{\delta\beta_1}{\delta(Y-W)}(S)\tag{3-5-12}$$

常数律：
$$\frac{\delta}{\delta Y}K=0,\quad \text{如果 } Y\neq\varnothing\tag{3-5-13}$$

其中 K 为常数。

幂律：如果 $p(S)$ 是一个置信函数，它具有密度函数 $f_p(\boldsymbol{y})$，则

$$\frac{\delta}{\delta Y}p(S)^n=\begin{cases}\dfrac{n!}{(n-k)!}p(S)^{n-k}f_p(\boldsymbol{y}_1)\cdots f_p(\boldsymbol{y}_k), & k\leqslant n\\ 0, & k>n\end{cases}\tag{3-5-14}$$

例如，对于量测集 S 和状态集 X，S 关于 X 的条件置信函数如下

$$\begin{cases}\beta_k(S\mid X)=P(\Sigma_k\subseteq S)=\int_S f_k(Z\mid X)\delta Z\\ \beta_{k+1|k}(S\mid X)=P(\Sigma_{k+1|k}\subseteq S)=\int_S f_{k+1|k}(Y\mid X)\delta Y\end{cases}\tag{3-5-15}$$

其中 Σ_k,Z 均是 RFS，那么可以通过求导获得如下的密度函数

$$\begin{cases}f_k(Z\mid X)=\dfrac{\delta\beta_k}{\delta Z}(\varnothing\mid X)\\ f_{k+1|k}(Y\mid X)=\dfrac{\delta\beta_{k+1|k}}{\delta Y}(\varnothing\mid X)\end{cases}\tag{3-5-16}$$

例题 3.5.2 考虑杂波条件下单目标似然函数。假设在 k 时刻，Σ_k 为量测 RFS，X_k 为目标状态集，其中目标量测 RFS 为 Φ_k，杂波量测 RFS 为 C_k，即 $\Sigma_k=\Phi_k\cup C_k$，量测数据集为 $Y_k=\{y_k^i\}$。那么，在量测空间 S_o 中置信函数为

$$\beta_{\Sigma_k}(S_o \mid X_k)=P(\Sigma_k \subseteq S_o \mid X_k)=P(\Phi_k \subseteq S_o \mid X_k)P(C_k \subseteq S_o)$$
$$=\beta_{\Phi_k}(S_o \mid X_k)\beta_{C_k}(S_o)$$

利用乘积律，对上式求导并取空集

$$\frac{\delta}{\delta Y_k}(\beta_{\Phi_k}(S_o \mid X_k)\beta_{C_k}(S_o))=\sum_{W\subseteq Y_k}\frac{\delta\beta_{\Phi_k}}{\delta W}(S)\frac{\delta\beta_{C_k}}{\delta(Y_k-W)}(S)$$
$$=\sum_{Z_k\subseteq Y_k}\frac{\delta\beta_{\Phi_k}}{\delta Z_k}(S)\frac{\delta\beta_{C_k}}{\delta(Y_k-Z_k)}(S)$$
$$=\sum_{Z_k\subseteq Y_k}f_{\Phi_k}(Z_k \mid X_k)f_{C_k}(Y_k-Z_k)$$
$$=p_c(m_k)(1-P_D)+p_c(m_k-1)P_D\sum_i f_{\Phi_k}(z_k^i \mid x_k) \quad \square$$

例题 3.5.3 两个目标似然函数，对于如下的置信函数

$$\beta_\Sigma(S \mid \{\boldsymbol{x}\})=1-P_D+P_D p_f(S \mid \boldsymbol{x})$$
$$\beta_\Sigma(S \mid \{\boldsymbol{x}_1,\boldsymbol{x}_2\})=[1-P_D+P_D p_f(S \mid \boldsymbol{x}_1)][1-P_D+P_D p_f(S \mid \boldsymbol{x}_2)]$$

那么，似然密度分别为

$$f(\varnothing \mid \varnothing)=1$$
$$f(\varnothing \mid \{\boldsymbol{x}\})=1-P_D$$
$$f(\{\boldsymbol{z}\} \mid \{\boldsymbol{x}\})=P_D f(\boldsymbol{z} \mid \boldsymbol{x})$$
$$f(\varnothing \mid \{\boldsymbol{x}_1,\boldsymbol{x}_2\})=(1-P_D)^2$$
$$f(\{\boldsymbol{z}\} \mid \{\boldsymbol{x}_1,\boldsymbol{x}_2\})=P_D(1-P_D)f(\boldsymbol{z} \mid \boldsymbol{x}_1)+P_D(1-P_D)f(\boldsymbol{z} \mid \boldsymbol{x}_2)$$
$$f(\{\boldsymbol{z}_1,\boldsymbol{z}_2\} \mid \{\boldsymbol{x}_1,\boldsymbol{x}_2\})=P_D^2 f(\boldsymbol{z}_1 \mid \boldsymbol{x}_1)f(\boldsymbol{z}_2 \mid \boldsymbol{x}_2)+P_D^2 f(\boldsymbol{z}_2 \mid \boldsymbol{x}_1)f(\boldsymbol{z}_1 \mid \boldsymbol{x}_2)$$

其中，$\boldsymbol{x},\boldsymbol{x}_1,\boldsymbol{x}_2$ 为不同目标的状态，$\boldsymbol{z},\boldsymbol{z}_1,\boldsymbol{z}_2$ 对应各自的量测；$p_f(S|\boldsymbol{x})$是密度为 $f(\boldsymbol{z}|\boldsymbol{x})$ 的概率分布函数。 $\square$

3.5.4 典型 RFS（随机有限集）分布函数

分布函数是随机有限集理论的基础，依据分布函数的不同，典型的 RFS 包括 Poisson RFS、单 Bernoulli RFS、多 Bernoulli RFS 等。下面对几种 RFS 做一个简要的介绍。

1. Poisson RFS

对于状态空间$\mathbb{X}$ 上的 RFS X，若其势$|X|$满足 Poisson 分布，就称为 Poisson RFS，用 $v(\boldsymbol{x})$表示其强度函数，则其均值和概率密度函数可以分别表示为 $\bar{N}=\int v(\boldsymbol{x})\mathrm{d}\boldsymbol{x}$ 和 $v(\boldsymbol{x})/\bar{N}$。Poisson RFS 的概率密度函数可以表示为

$$\pi(X)=\mathrm{e}^{-\bar{N}}v^X \tag{3-5-17}$$

对于状态空间上的 $h:\mathbb{X}\to\mathbb{R}$，满足 $h^X=\prod_{x\in X}h(\boldsymbol{x})$，规定 $h^\varnothing=1$。

2. Bernoulli RFS

对于状态空间$\mathbb{X}$ 上的 Bernoulli RFS，若用 p 表示其元素分布的概率密度函数，用 r

表示其存在概率，则其概率密度函数可以表示为

$$\pi(X)=\begin{cases}1-r, & X=\varnothing\\ r\cdot p(\boldsymbol{x}), & X=\{\boldsymbol{x}\}\end{cases} \tag{3-5-18}$$

3. 多 Bernoulli RFS

对于多 Bernoulli RFS，它是个数固定且彼此之间相互独立的 Bernoulli RFS 的集合，分别用 $r^{(i)}$，$p^{(i)}$ 表示每个 Bernoulli RFS 的存在概率和概率密度函数，且满足 $r^{(i)}\in(0,1)$，$i=1,2,\cdots,M$，$X=\bigcup_{i=1}^{M}X^{(i)}$，$M$ 表示其中所包含的 Bernoulli RFS 的数量。由于 Bernoulli RFS 的概率密度函数仅仅与参数 r 和 p 有关，所以完全可以用参数集 $\{(r^{(i)},p^{(i)})\}_{i=1}^{M}$ 对其表示。对参数集进行预测更新即可，其势分布可以表示为 $\sum_{i=1}^{M}r^{(i)}$，概率密度函数如下

$$\begin{cases}\pi(\phi)=\prod_{j=1}^{M}(1-r^{(j)})\\ \pi(\{\boldsymbol{x}_1,\cdots,\boldsymbol{x}_n\})=\pi(\phi)\sum_{\{i_1,\cdots,i_n\}\in\mathcal{F}_n(\{1,\cdots,M\})}\prod_{j=1}^{n}\dfrac{r^{(i_j)}p^{(i_j)}(\boldsymbol{x}_j)}{1-r^{(i_j)}}\end{cases} \tag{3-5-19}$$

3.5.5 标签 RFS

不同于传统的 RFS，标签 RFS 为不同目标的状态 $\boldsymbol{x}(\boldsymbol{x}\in\mathbb{X})$ 添加了可以识别身份的标签 $l\in\mathbb{L}=\{\alpha_i:i\in N\}$，其中 N 表示目标的个数。为区别非标签 RFS 变量，多目标的状态用粗体标签 RFS 表示为[94-95]

$$\boldsymbol{X}_k=\{(\boldsymbol{x}_{k,1},l_1),\cdots,(\boldsymbol{x}_{k,N(k)},l_{N(k)})\} \tag{3-5-20}$$

其中用大写实体的 $\boldsymbol{X}_k$ 表示多目标状态集合，用小写实体 $\boldsymbol{x}_{k,i}$ 表示目标 i 在 k 时刻的带标签状态。为了实现不同目标的标签是不同的，用下述方程进行约束

$$\Delta(\boldsymbol{X})=\begin{cases}1, & |\mathcal{L}(\boldsymbol{X})|=|\boldsymbol{X}|\\ 0, & |\mathcal{L}(\boldsymbol{X})|\neq|\boldsymbol{X}|\end{cases} \tag{3-5-21}$$

其中，$\mathcal{L}(\boldsymbol{X})$ 表示状态集到 $\boldsymbol{X}$ 标签 l 的映射，$|\cdot|$ 表示集合中元素的个数。

简单地说，标签随机有限集(L-RFS)就是在标准 RFS 中为每个元素增添一个独一无二的标签。这样一来可以用来识别定位跟踪区域内的每个目标，从而生成有效航迹。用 $\mathbb{L}=\{\alpha_i:i\in N\}$ 表示标签空间，在原来的状态 $\boldsymbol{x}\in\mathbb{X}$ 上添加一个标签信息 $l\in\mathbb{L}$，则目标状态 $(\boldsymbol{x},l)$ 定义在了空间 $\mathbb{X}\times\mathbb{L}$ 上。

1. 标签 Poisson RFS

类比于前面提及的 Poisson RFS，标签泊松 RFS 只是添加了一个标签信息，这里就不做过多赘述，直接给出概率密度函数的表达式[95]

$$\pi(\{(\boldsymbol{x}_1,l_1),\cdots,(\boldsymbol{x}_n,l_n)\})=\delta_{L(n)}(\{l_1,\cdots,l_n\})\mathrm{Pois}_{\langle v,1\rangle}(n)\prod_{i=1}^{n}\frac{v(\boldsymbol{x}_i)}{\langle v,1\rangle} \tag{3-5-22}$$

其中，$\mathrm{Pois}_{\lambda}(n)=\mathrm{e}^{-\lambda}\lambda^{n}/n!$ 表示参数为 λ 的 Poisson 分布，$L(n)=\{\alpha_i\in L\}_{i=1}^{n}$。

2. 标签多 Bernoulli RFS

若将标签多 Bernoulli RFS 的参数集表示为$\{(r^{(\zeta)},p^{(\zeta)}):\zeta\in\Psi\}$，将与之对应的标签状态表示为 $\alpha(\zeta)$，其中，$\alpha:\Psi\to L$ 是一个单射，在原来状态信息的基础上增添了标签信息，把原来状态空间$\mathbb{X}$ 上的问题扩展到了空间$\mathbb{X}\times\mathbb{L}$ 上，其概率密度函数可以表示为[95]

$$\pi(\{(\boldsymbol{x}_1,l_1),\cdots,(\boldsymbol{x}_n,l_n)\})=\delta_{(n)}(|\{l_1,\cdots,l_n\}|)\times\prod_{\zeta\in\Psi}(1-r^{(\zeta)})\prod_{j=1}^{n}\frac{1_{\alpha(\Psi)}(l_j)r^{(\alpha^{-1}(l_j))}p^{(\alpha^{-1}(l_j))}(\boldsymbol{x}_j)}{1-r^{(\alpha^{-1}(l_j))}} \tag{3-5-23}$$

简化公式为

$$\pi(\boldsymbol{X})=\Delta(\boldsymbol{X})1_{\alpha(\Psi)}(\mathcal{L}(\boldsymbol{X}))[\Phi(X;\cdot)]^{\Psi} \tag{3-5-24}$$

其中，$\Delta(\boldsymbol{X})$为表示标签是否唯一的一个指标，$\delta_{|X|}(|\mathcal{L}(\boldsymbol{X})|)$为指示函数，用来表示是否恰好为定位跟踪区域内的每个目标添加了一个唯一的标签，$\mathcal{L}(\boldsymbol{X})$表示 $\boldsymbol{X}$ 的标签，则有

$$\Phi(\boldsymbol{X};\zeta)=\begin{cases}1-r^{(\zeta)}, & \alpha(\zeta)\notin\mathcal{L}(\boldsymbol{X})\\ r^{(\zeta)}p^{(\zeta)}(\boldsymbol{x}), & (\boldsymbol{x},\alpha(\zeta))\in\boldsymbol{X}\end{cases} \tag{3-5-25}$$

3. 广义标签多 Bernoulli RFS

GLMB 为空间$\mathbb{X}\times\mathbb{L}$ 上的标签 RFS，其分布为[95]

$$\pi(\boldsymbol{X})=\Delta(\boldsymbol{X})\sum_{c\in C}w^{(c)}(\mathcal{L}(\boldsymbol{X}))[p^{(c)}]^{X} \tag{3-5-26}$$

其中，C 为离散变量，$w^{(c)}(L)$，$p^{(c)}$分别满足

$$\begin{cases}\sum\limits_{l\in L}\sum\limits_{c\in C}w^{(c)}(l)=1\\ \int p^{(\zeta)}(\boldsymbol{x},l)\mathrm{d}\boldsymbol{x}=1\end{cases} \tag{3-5-27}$$

3.5.6 随机有限集的 Bayes 滤波

考虑传统的 Bayes 滤波公式为

$$p_{k|k-1}(\boldsymbol{x}_k\mid\boldsymbol{z}_{1:k-1})=\int p_{k|k-1}(\boldsymbol{x}_k\mid\boldsymbol{x}_{k-1})p_{k-1|k-1}(\boldsymbol{x}_{k-1}\mid\boldsymbol{z}_{1:k-1})\mathrm{d}\boldsymbol{x}_{k-1} \tag{3-5-28}$$

$$p_{k|k}(\boldsymbol{x}_k\mid\boldsymbol{z}_{1:k})=\frac{g_k(\boldsymbol{z}_k\mid\boldsymbol{x}_k)p_{k|k-1}(\boldsymbol{x}_k\mid\boldsymbol{z}_{1:k-1})}{\int p_k(\boldsymbol{z}_k\mid\boldsymbol{x}_k)p_{k|k-1}(\boldsymbol{x}_k\mid\boldsymbol{z}_{1:k})\mathrm{d}\boldsymbol{x}_k} \tag{3-5-29}$$

其中 $\boldsymbol{z}_{1:k-1}\triangleq\{\boldsymbol{z}_1,\boldsymbol{z}_2,\cdots,\boldsymbol{z}_{k-1}\}$。类似地，对于随机有限集来说，基于信度函数上的集合积分和导数，Mahler 给出了 FISST 框架下的 Bayes 公式如下

$$f_{k|k-1}(X_k \mid Z_{1:k-1}) = \int f_{k|k-1}(X_k \mid X_{k-1}) f_{k-1|k-1}(X_{k-1} \mid Z_{1:k-1}) \delta X_{k-1} \tag{3-5-30}$$

$$f_{k|k}(X_k \mid Z_{1:k}) = \frac{\rho_k(Z_k \mid X_k) f_{k|k-1}(X_k \mid Z_{1:k-1})}{\int \rho_k(Z_k \mid X_k) f_{k|k-1}(X_k \mid Z_{1:k-1}) \delta X_k} \tag{3-5-31}$$

其中 $X_k, Z_{1:k-1}$ 分别是系统状态和直到 $k-1$ 时刻的量测；而 $f_{k|k}(\cdot), f_{k|k-1}(\cdot), \rho_k(\cdot)$ 分别是状态更新、状态提前一步预测和量测噪声的置信密度函数。

FISST 框架下的 Bayes 公式和传统概率框架下的 Bayes 公式有何异同？有何联系？Vo 等给出了如下的命题[65]。

命题 1 假设 Σ 是局部紧的 Hausdorff 可分空间 E 上的 RFS，并且具有概率测度 P_Σ 和信度函数 β_Σ，如果 P_Σ 是关于 μ 绝对连续，μ 是一个没有进行正则化的 Poisson 点过程测度，具有单位 K^{-1}，对于任意子集 $A \in \mathcal{F}(E)$，定义如下 Poisson 点过程测度

$$\mu(A) = \sum_{i=0}^{\infty} \frac{\lambda^i (\Sigma^{-1}(A) \cap E^i)}{i!}$$

那么

$$\frac{\mathrm{d}P_\Sigma}{\mathrm{d}\mu} = K^{|X|} d(\beta_\Sigma)_X \tag{3-5-32}$$

这说明 FISST 框架下的置信密度函数是带有单位 $K^{-|X|}$ 的，传统的 Bayes 公式和 FISST 框架下的 Bayes 公式相差单位 $K^{-|X|}$。因此，二者之间存在如下关系

$$\begin{cases} p_{k-1|k-1}(X \mid Z_{1:k-1}) = K_s^{|X|} f_{k-1|k-1}(X \mid Z_{1:k-1}) \\ p_{k|k-1}(X \mid Z_{1:k-1}) = K_s^{|X|} f_{k|k-1}(X \mid Z_{1:k-1}) \\ p_{k|k-1}(X \mid Z_{1:k-1}) = K_s^{|X|} f_{k|k-1}(X \mid Z_{1:k-1}) \\ p_{k|k-1}(X \mid X_{k-1}) = K_s^{|X|} f_{k|k-1}(X \mid X_{k-1}) \\ p_{k|k-1}(X \mid X_{k-1}) = K_s^{|X|} f_{k|k-1}(X \mid X_{k-1}) \\ p_{k|k-1}(X \mid X_{k-1}) = K_s^{|X|} f_{k|k-1}(X \mid X_{k-1}) \\ p_{k|k-1}(X \mid X_{k-1}) = K_s^{|X|} f_{k|k-1}(X \mid X_{k-1}) \\ p_{k|k-1}(X \mid X_{k-1}) = K_s^{|X|} f_{k|k-1}(X \mid X_{k-1}) \\ p_{k|k}(X \mid Z_{1:k}) = K_s^{|X|} f_{k|k}(X \mid Z_{1:k}) \\ g_k(Z \mid X_k) = K_o^{|Z|} \rho_k(Z \mid X_k) \end{cases} \tag{3-5-33}$$

其中，$p_{k|k}(\cdot), p_{k|k-1}(\cdot)$ 分别表示通常概率意义下的先验密度函数和提前一步预测密度函数，$g_k(\cdot)$ 是似然函数；而 $f_{k|k}(\cdot), f_{k|k-1}(\cdot), \rho_k(\cdot)$ 分别是相应的置信密度函数。

对随机变量(向量)而言，在线性高斯情况下，利用标准的 Kalman 滤波器，就可以更新状态变量的一阶矩和二阶矩的估计。然而，对于随机有限集而言，是否存在只需更新前两阶矩的滤波公式呢？这就需要解决两个问题：首先要确定随机有限集的矩是否存在；其次在矩存在的前提下，是否存在其递推公式。此外，随机有限集 Bayes 公式的计算也有着本质的困难。

(1) 计算难度大大增加。根据式(3-5-29)的随机有限集 Bayes 公式中包含一系列的

高维积分，这个在实际中很难应用。此外，式(3-5-29)存在一个假设：置信密度函数 $f(\{\boldsymbol{y}_1,\cdots,\boldsymbol{y}_n\})$的分布是对称的，即对于序列$\{\boldsymbol{y}_1,\cdots,\boldsymbol{y}_n\}$的任意的序列重排$\{\boldsymbol{y}_{i_1},\cdots,\boldsymbol{y}_{i_n}\}$，$f(\{\boldsymbol{y}_{i_1},\cdots,\boldsymbol{y}_{i_n}\})$与 $f(\{\boldsymbol{y}_1,\cdots,\boldsymbol{y}_n\})$二者具有相同的分布。

(2) 估计准则的含义是什么？是否存在？对于传统的随机变量(向量)，常用的Bayes估计准则包括最小均方差估计(MSE)和极大后验估计(MAP)，分别定义如下：

$$\text{MSE：}\ \hat{\boldsymbol{x}}_k = E(\boldsymbol{x}_k \mid \boldsymbol{z}_{1:k})$$

$$\text{MAP：}\ \hat{\boldsymbol{x}}_k = \arg\max_{x_k} f(\boldsymbol{x}_k \mid \boldsymbol{z}_{1:k})$$

但是，对于随机有限集 X_k 来说，条件期望 $E(X_k \mid Z_{1:k})$没有任何意义，或者说根本不存在，更谈不上 MSE 准则问题。MAP 估计则与状态的单位有关系。因此，这些估计准则均不适用于随机有限集的估计。这些困难促使人们寻找更为可行、简化和实用的算法，这一点，我们将在随机有限集多目标跟踪理论中进一步讨论。

3.6 统计学习理论与支持向量机基础

1963 年 Vapnik 在解决模式识别问题时提出了支持向量方法，这种方法从训练集中选择一组特征子集，使得对特征子集的划分等价于对整个数据集的划分，这组特征子集被称为支持向量(support vector，SV)，这种理论称为支持向量机(support vector machine，SVM)理论。1971 年 Kimeldorf 提出使用线性不等约束重新构造 SV 的核空间，解决了一部分线性不可分问题；1990 年，Grace、Boser 和 Vapnik 等开始对 SVM 进行系统研究；1995 年 Vapnik 正式提出统计学习理论[70]。SVM 理论越来越受到人们的关注，而且已经几乎成为数据分类的一种标准方法。

3.6.1 统计学习理论的一般概念

机器学习是一种基于数据的学习方法，它主要研究从观测数据(样本)出发寻找规律并构造一个模型，利用该模型可对未知数据或者无法观测的数据进行预测。这种模型被称为**学习机(learning machine)**。机器学习的过程就是构建学习机的过程。机器学习包含很多种方法，如决策树、遗传算法、人工神经网络、隐 Markov 链(HMM)等。然而迄今为止，关于机器学习还没有一种被共同接受的理论框架，但关于实现方法大致可以分为三种。

第一种是经典的(参数)**统计估计方法(statistical estimation method)**。在这种方法下，参数的依赖关系被认为是先验已知的，训练样本用于估计依赖关系的参数。这种方法的缺点在于：一是因为“维数灾难”，很难将这种方法扩展到高维数据空间；二是先验的依赖关系在很多情况下难于获取。

第二种方法是 20 世纪 80 年代发展起来的**经验非线性方法(empirical nonlinear method)**，如人工神经网络(ANN)和多变量自适应回归分析。这些方法利用已知样本建立非线性模型，克服了传统参数估计方法的困难。但是，这些方法目前还缺乏统一的数学理论指导。

第三种方法是20世纪60年代后期发展起来的**统计学习理论**(**statistical learning theory 或 SLT**),它是专门用于有限样本情况下非参数估计问题的机器学习理论。与传统统计学相比,这种体系下的统计推理规则不仅考虑了对渐近性能的要求,而且追求在现有有限信息的条件下得到最优结果。Vapnik等从20世纪60至70年代开始致力于此方面研究,到90年代中期其理论发展已趋于成熟,而且统计学习理论开始受到越来越广泛的重视。

统计学习理论建立在一套较坚实的理论基础之上,为解决有限样本学习问题提供了一个统一的框架。它能将很多现有方法纳入其中,有望帮助解决许多原来难以解决的问题(如神经网络结构选择问题、局部极小点问题等);同时,在这一理论基础上发展了一种新的通用学习方法——支持向量机理论,它已初步表现出很多优于已有方法的性能。一些学者认为,SLT和SVM正在成为继神经网络研究之后新的研究热点,并将有力地推动机器学习理论和技术的发展。

机器学习的目的是根据给定的训练样本求取系统输入输出之间依赖关系的估计,使它能够对系统行为做出尽可能准确的预测。

定义 3.6.1 设变量 y 与变量 $\boldsymbol{x}$ 之间存在一定的未知依赖关系,即遵循某一未知的联合概率分布 $P(\boldsymbol{x},y)$,机器学习问题就是根据 l 个独立同分布(i.i.d)的观测样本

$$(\boldsymbol{x}_1,y_1),(\boldsymbol{x}_2,y_2),\cdots,(\boldsymbol{x}_l,y_l) \tag{3-6-1}$$

在一组函数$\{f(\boldsymbol{x},\alpha)\}$中按某个准则选择一个最优的函数 $f(\boldsymbol{x},\alpha_0)$对依赖关系进行估计,其中 α 是参数,而 $f(\boldsymbol{x},\alpha)$就称为是一个**学习机**(**learning machine**)。如果对于给定的输入 $\boldsymbol{x}$ 和 α 的一个选择,其输出 $f(\boldsymbol{x},\alpha)$总是相同的,这个学习机称为是**确定性的**(**deterministic**);否则称为**不确定性的**(**uncertainty**)。对于确定性学习机中 α 的一个特定选择,就称其为**训练过的学习机**(**trained machine**)。

例题 3.6.1 考虑具有固定结构的一个人工神经网络,α 就是所有的权系数和偏置,在此意义下就是一个学习机。如果给定 α 的一组值,这个人工神经网络就是一个训练过的学习机。显然,给定一组输入 $\boldsymbol{x}$ 和 α 的值,这个人工神经网络输出为确定值,则其为确定性的人工神经网络。 □

定义 3.6.2 对于一个学习机,**损失函数**(**loss function**)$L(y,f(\boldsymbol{x},\alpha))$表示用 $f(\boldsymbol{x},\alpha)$对 y 进行预测而造成的损失。不同类型的学习问题有不同形式的损失函数,针对三类主要的学习问题分别定义如下。

(1) **模式识别问题**:输出变量 y 是类别标号;对于只有两种模式的情况 $y=\{0,1\}$或者 $y=\{-1,1\}$,此时损失函数一般定义为

$$L(y,f(\boldsymbol{x},\alpha))=\begin{cases}0(\text{或}-1), & y=f(\boldsymbol{x},\alpha)\\ 1, & y\neq f(\boldsymbol{x},\alpha)\end{cases} \tag{3-6-2}$$

(2) **函数逼近问题**:$y\in\mathbb{R}$ 是连续变量,损失函数一般以定义为

$$L(y,f(\boldsymbol{x},\alpha))=(y-f(\boldsymbol{x},\alpha))^2 \quad \text{或} \quad L(y,f(\boldsymbol{x},\alpha))=\frac{1}{2}\,|\,y-f(\boldsymbol{x},\alpha)\,| \tag{3-6-3}$$

(3) **概率密度估计问题**:学习的目的是根据训练样本决定 x 的概率密度,设被估计的概率密度为 $p(\boldsymbol{x},\alpha)$,则损失函数一般定义为

$$L(p(\boldsymbol{x},\alpha)) = -\log p(\boldsymbol{x},\alpha) \tag{3-6-4}$$

定义 3.6.3 对于一个学习机，损失函数的期望

$$R(\alpha) = \int L(y, f(\boldsymbol{x},\alpha))\mathrm{d}P(\boldsymbol{x},y) \tag{3-6-5}$$

称为**期望风险**（**expected risk**）或**真实风险**（**actual risk**）。而**经验风险**（**empirical risk**）则是训练集上的平均误差，即

$$R_{\text{emp}}(\alpha) = \frac{1}{l}\sum_{i=1}^{l} L(y_i, f(\boldsymbol{x}_i,\alpha)) \tag{3-6-6}$$

注意经验风险中并没有出现概率分布。对于特定的 α 的一个选择，以及特定的训练样本 $(\boldsymbol{x}_i, y_i)$，经验风险是一个确定的值。

传统的机器学习方法采用了所谓的**经验风险最小化**（ERM）准则，即用样本定义经验风险如式（3-6-6）所示。该式作为对式（3-6-5）的估计，设计学习算法使其最小化，这就是经验风险最小化准则。

3.6.2 学习机的 VC 维与风险界

定义 3.6.4 考虑相应于两个模式类的识别问题，其中函数 $f(\boldsymbol{x},\alpha)\in\{1,-1\}$，$\forall \boldsymbol{x}$，$\alpha$。对于给定的一组 l 个点，可以按所有 2^l 个可能的方式进行标识；对于每个可能的标识，如果在 $\{f(\alpha)\}$ 中总能取得一个函数，这个函数能够用来正确地区分这些标识，称这 l 个点构成的点集能够被 $\{f(\alpha)\}$ **分隔**（**shattered**）。函数族 $\{f(\alpha)\}$ 的 **VC 维**（**Vapnik Chervonenkis dimension**）定义为能够被 $\{f(\alpha)\}$ 分隔的训练点 l 的最大数目。若对任意数目的样本都有函数能将它们分隔，则函数集的 VC 维是无限大。

注意，如果 VC 维是 h，那么至少存在一组 h 个点的集合能够被分隔。但是在一般情况下，任意一组 h 个点的集合都能被分隔的结论却是不正确的。

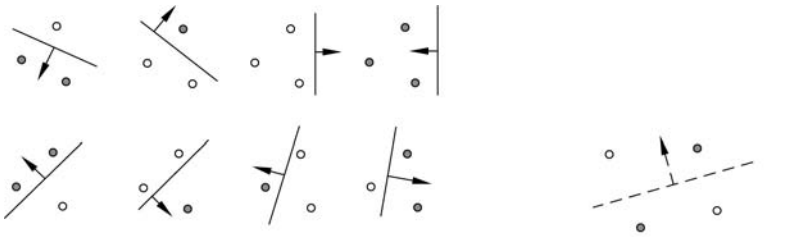
例题 3.6.2 考虑一个分类问题。假定所有的数据来自空间 $\mathbb{R}^2$。对于空间的任意三个点的集合 $\{\boldsymbol{x}_1,\boldsymbol{x}_2,\boldsymbol{x}_3\}$，分为“左”“上”“下”，而且有 8 种（$2^3$）可能的标识（见图 3-6-1(a)）

$$\begin{cases} \boldsymbol{x}_1\to 1,\boldsymbol{x}_2\to -1,\boldsymbol{x}_3\to 1; & \boldsymbol{x}_1\to -1,\boldsymbol{x}_2\to 1,\boldsymbol{x}_3\to -1; & \boldsymbol{x}_1\to -1,\boldsymbol{x}_2\to -1,\boldsymbol{x}_3\to -1 \\ \boldsymbol{x}_1\to 1,\boldsymbol{x}_2\to 1,\boldsymbol{x}_3\to 1; & \boldsymbol{x}_1\to 1,\boldsymbol{x}_2\to 1,\boldsymbol{x}_3\to -1; & \boldsymbol{x}_1\to -1,\boldsymbol{x}_2\to -1,\boldsymbol{x}_3\to 1 \\ \boldsymbol{x}_1\to 1,\boldsymbol{x}_2\to -1,\boldsymbol{x}_3\to -1; & \boldsymbol{x}_1\to -1,\boldsymbol{x}_2\to 1,\boldsymbol{x}_3\to 1 \end{cases}$$

对于这种标识，总可以找到一组有向直线将不同的标识点进行分隔（注意标识不是赋值）。

$\{f(\alpha)\}$ 由有向直线组成，对于给定的直线，在直线一边的所有点规定属于类别 1，而所有在另一边的点规定属于类别 −1。直线的方向在图 3-6-1(a)中用箭头表示，其正方向一边的点标识为 1。所以 $\mathbb{R}^2$ 中一组有向直线的 VC 维是 3。然而，同一直线上的三个点任意标识，却不能保证被有向直线分隔，这却不影响 $\mathbb{R}^2$ 中一组有向直线的 VC 维是 3。

但是，对于空间四个点的集合 $\{\boldsymbol{x}_1,\boldsymbol{x}_2,\boldsymbol{x}_3,\boldsymbol{x}_4\}$，分为“左上”“右上”“左下”“右下”，我们仅取一种标识（见图 3-6-1(b)）：$\boldsymbol{x}_1\to -1,\boldsymbol{x}_2\to 1,\boldsymbol{x}_3\to 1,\boldsymbol{x}_4\to -1$，任何一个有向直线都不能把不同的标识点进行分隔。其实，对于任意四个点的集合，采用 16 种可能的标识，都不可能找到一组有向直线将不同的标识点完全进行分隔。 □



(a) 三个点被有向直线分隔　　(b) 四个点不能被有向直线分隔

图 3-6-1　$\mathbb{R}^2$ 中的 VC 维

现在考虑$\mathbb{R}^n$ 中的有向超平面，下面的定理将非常有用。

引理 1　$\mathbb{R}^n$ 中两个点集能够被一个超平面分隔的充分与必要条件是其凸包(包含该点集的最小凸集)的交集为空集。

证明　见文献[1,67]。 ■

定理 3.6.1　考虑$\mathbb{R}^n$ 中 m 个点的集合，选择其中任何一个点作为原点，那么 m 个点能够被有向超平面分隔的充分与必要条件是，除了原点外其余点的位置向量是线性独立的。

证明

(1) 其余点的位置向量线性独立⇒m 个点能够被有向超平面分隔。把原点标识为 O，而且假定其余的 $m-1$ 个点的位置向量是线性独立的。考虑对任意一种可能的标识，按二值标识把 m 个点划分为两个子集 S_1 和 S_2，分别有 m_1 和 m_2 个点，所以 $m_1+m_2=m$。不妨设 S_1 包含点 O，则根据凸组合的定义，S_1 的凸包 C_1 中所有点的位置向量 $\boldsymbol{x}$ 满足

$$\boldsymbol{x}=\sum_{i=1}^{m_1}\alpha_i\boldsymbol{s}_{1i},\quad \sum_{i=1}^{m_1}\alpha_i=1,\quad \alpha_i\geqslant 0 \tag{3-6-7}$$

其中 $\boldsymbol{s}_{1i}$ 就是 S_1 中 m_1 个点的位置向量(包括原点的零向量)。类似地，S_2 的凸包 C_2 中所有点的位置向量 $\boldsymbol{x}$ 满足

$$\boldsymbol{x}=\sum_{i=1}^{m_2}\beta_i\boldsymbol{s}_{2i},\quad \sum_{i=1}^{m_2}\beta_i=1,\quad \beta_i\geqslant 0 \tag{3-6-8}$$

其中 $\boldsymbol{s}_{2i}$ 就是 S_2 中 m_2 个点的位置向量。

现在假定 $C_1\cap C_2=D\neq\varnothing$，存在 $\boldsymbol{x}\in D\subseteq\mathbb{R}^n$ 同时满足式(3-6-4)和式(3-6-5)，两方程相减得

$$\sum_{i=1}^{m_1}\alpha_i\boldsymbol{s}_{1i}-\sum_{j=1}^{m_2}\alpha_j\boldsymbol{s}_{2j}=0 \tag{3-6-9}$$

这就表明这两组 $m-1$ 个非零位置向量线性相关，这与线性独立的假设相矛盾。再根据引理 1，因为 $C_1\cap C_2=\varnothing$，则存在一个超平面分隔 S_1 和 S_2。

(2) m 个点能够被有向超平面分隔⇒其余点的位置向量线性独立。反设 $m-1$ 个非零位置向量线性相关，则存在 $m-1$ 个数 γ_i 使得

$$\sum_{i=1}^{m-1}\gamma_i\boldsymbol{s}_i=0 \tag{3-6-10}$$

如果所有的 γ_i 具有相同的符号，则可以通过比例变换使得 $\gamma_i \in [0,1]$，且 $\sum_i \gamma_i = 1$。

上式表明原点处在其他点的凸包中，根据引理 1，不存在超平面可以把原点与其他点分离，所以 m 个点不能够被有向超平面分隔。

如果所有的 γ_i 不具有相同的符号，把所有负号的放在一边，得

$$\sum_{i \in I_1} |\gamma_i| \boldsymbol{s}_i = \sum_{j \in I_2} |\gamma_j| \boldsymbol{s}_j \tag{3-6-11}$$

其中 I_1, I_2 是对集合 $S-O$ 的划分所对应的指标集合。变换比例使得下面两组中的任意一组成立

$$\begin{cases} \sum_{i \in I_1} |\gamma_i| = 1, & \text{且} \sum_{j \in I_2} |\gamma_j| \leqslant 1 \\ \sum_{i \in I_1} |\gamma_i| \leqslant 1, & \text{且} \sum_{j \in I_2} |\gamma_j| = 1 \end{cases} \tag{3-6-12}$$

不失一般性，假设后者成立。那么方程(3-6-7)的左边是点集 $\left\{ \sum_{i \in I_1} \boldsymbol{s}_i \right\} \cup O$ 的凸包中点的位置向量，右边就是点集 $\sum_{j \in I_2} \boldsymbol{s}_j$ 的凸包中点的位置向量，所以凸包的交集不空。根据引理 1，不存在超平面可以把两个凸包分离，所以 m 个点不能够被有向超平面分隔。

这与假设相矛盾，所以结论成立。 ■

推论 $\mathbb{R}^n$ 中有向超平面集合的 VC 维是 $n+1$。

证明 因为我们总能在 $\mathbb{R}^n$ 中选择 $n+1$ 个点，其中一个点作为原点，而其余 n 个点的位置向量线性独立；但是，我们却不能在 $\mathbb{R}^n$ 中选择 $n+2$ 个点，其中一个点作为原点，而其余 $n+1$ 个点的位置向量线性独立（$\mathbb{R}^n$ 维数的限制），从而结论成立。 ■

注 在 $\mathbb{R}^n$ 中至少可以找到一组 $n+1$ 个点能够被超平面分隔，但不是任意一组 $n+1$ 个点都能够被超平面分隔。例如把所有点集中在一条直线上的情况就不能被分隔。

定义 3.6.5 对于一个学习机，假定损失函数以概率 $1-\eta$ 取值，其中 $0 \leqslant \eta \leqslant 1$，那么下面的误差界成立

$$R(\alpha) \leqslant R_{\text{emp}}(\alpha) + \sqrt{\frac{h(\log(2l/h)+1) - \log(\eta/4)}{l}} \tag{3-6-13}$$

其中 h 是 VC 维，是对上述容量概念的一个度量，上式右边称为**风险界（risk bound）**，又称为推广性的界。它由两部分组成，第一项是训练误差造成的经验风险，第二项是置信范围，而后者与学习机的 VC 维及训练样本数有关。

此式表明，在有限训练样本下，学习机的 VC 维越高（复杂性越高），则置信范围越大，导致真实风险与经验风险之差可能越大。这就是为什么片面追求经验风险最小化会导致出现过学习（over fitting）现象。机器学习过程不但要使经验风险最小，还要使 VC 维尽量小以缩小置信范围，才能取得较小的实际风险，对新的样本才有较好的泛化能力。

关于式(3-6-13)的风险界，有三点必须注意。一是它不依赖于概率分布 $P(\boldsymbol{x}, y)$，因为我们只是假定训练数据和测试数据是按分布 $P(\boldsymbol{x}, y)$ 抽取的独立同分布的数据；二是上式左边通常是无法计算的；三是只要我们知道了 h 的值，上式右边的值是很容易计算的。

VC维是对给定的一组函数容量概念的具体化。直观上说，人们可能期望有较多参数的学习机具有更高的VC维，而有较少参数的学习机具有更低的VC维。但是这种直观的认识却是不对的。例如，下面例题中的学习机只有一个参数，但它的VC维却是无限大。

例题 3.6.3 考虑阶跃函数 $\theta(z)$，$z\in\mathbb{R}$，$\theta(z)=1$，$\forall z>0$；$\theta(z)=-1$，$\forall z\leqslant 0$。再考虑一个参数的函数族，定义为

$$f(x,\alpha)\equiv\theta(\sin(\alpha x)),\quad x,\alpha\in\mathbb{R} \tag{3-6-14}$$

选择任意一个 $l\in\mathbb{N}$，寻求 l 个点能够被这一族函数分隔。选择点 $x_i=10^{-i}$，$i=1,2,\cdots,l$；规定标识 $y_1,y_2,\cdots,y_l$，$y_i\in\{-1,1\}$。于是，如果我们选择

$$\alpha=\left(1+\sum_{i=1}^{l}\frac{(1-y_i)10^i}{2}\right)\pi \tag{3-6-15}$$

那么 $f(\alpha)$就能给出这种标识。图3-6-2是 $l=4$ 的一种标识的图形表示。所以这个学习机的VC维是无限大。需要注意的是，虽然在本例题中对任意大的 l，总能找到 l 个点构成的点集，按任意标识都可以被选择的函数族分隔，但并不意味着任意 l 个点构成的点集都可以被选择的函数族分隔。例如取 $l=4$，$x_i=i$，$i=1,2,3,4$；相应的 $y_1,y_2,y_4=-1$，$y_3=1$，就不能被式(3-6-15)的函数族分隔。

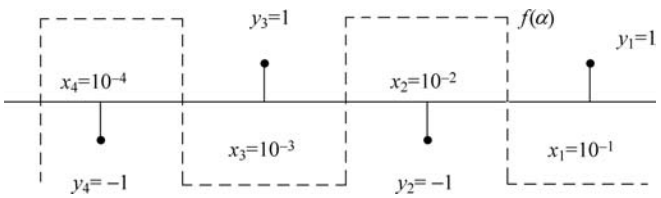


图 3-6-2 $l=4$ 的一种分隔 □

现在回过头来考察式(3-6-13)的界，其右边第二项随着 h 的变化而变化。给定一个置信水平95%($\eta=0.05$)，而且假定训练集的样本数是10000，VC置信范围将是 h 的单调增函数，无论 l 取值多大均是如此，如图3-6-3所示。于是，人们希望在经验风险为零的前提下，选择一个学习机，使得与其相关联的函数具有最小的VC维。然而，这将导致最差的误差界。于是，人们又希望在经验风险非零的前提下，选择一个学习机，使式(3-6-13)的右边为最小。如果采用这一策略，仅仅以式(3-6-13)为导向，只选择概率，就可得到实际风险的上界。这样做并不妨碍具有相同经验风险值的具体的学习机有很高的VC维，而得到更好的性能。

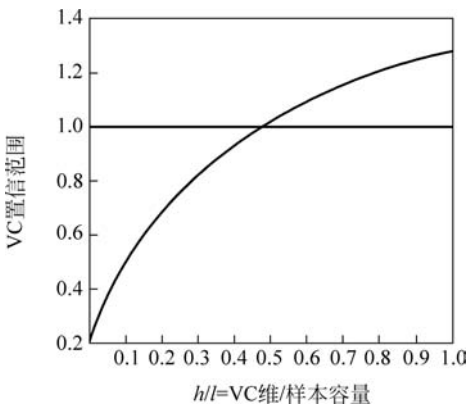


图 3-6-3 VC置信范围随 h 单调增

所以可以通过最小化 h 而达到最小化风险界的目的。

下面将讨论**结构风险最小化**(**structural risk minimization**，**SRM**)的问题。由式(3-6-13)

看出，要控制学习机器的实际风险，必须同时控制经验风险和置信范围。而为了使得式(3-6-13)右边两项同时最小化，必须使 VC 维成为一个可控制的变量。

注意，式(3-6-13)的 VC 置信范围依赖于函数类的选择，尽管经验风险和实际风险都通过训练过程依赖于具体函数的选择。我们希望在被选择的函数集中找到一些子集，使得对这个子集风险界达到最小。显然因为 VC 维是整数，我们难以使其平滑变化。于是，通过把整个函数族分解为嵌套的子集引入一种“结构”。对于每个子集，要么计算得到 VC 维 h 的值，要么得到 h 的界。于是，SRM 就是求取这些使得实际风险达到最小界的函数子集的一个过程。要做到这一点，通过简单地训练一组学习机来完成，一个学习机对应于一个子集；而对于给定的子集，训练的目的就是最小化经验风险。然后按序列取训练过的学习机，其经验风险与 VC 置信范围都是最小的。

统计学习理论提出的结构风险最小化方法，就是以经验风险和置信范围这两项最小化为目标的一种归纳方法。方法是把函数集$\{f(\boldsymbol{x},\alpha)\}$构造成一个嵌套的函数子集结构，即令 $S_n=\{f(\boldsymbol{x},\alpha)$：满足第 n 种约束$\}$，同时满足

$$S_1 \subset S_2 \subset \cdots \subset S_n \subset \cdots \tag{3-6-16}$$

各个子集对应的 VC 维满足

$$h_1 \leqslant h_2 \leqslant \cdots \leqslant h_n \leqslant \cdots \tag{3-6-17}$$

此时机器学习的任务是在每个子集中寻找经验风险最小的函数，在子集间折中考虑经验风险和置信范围，使得实际风险最小，如图 3-6-4 所示。

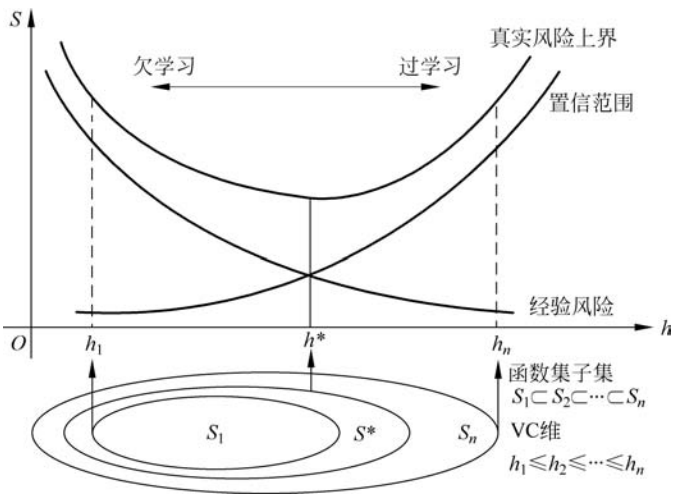


图 3-6-4　按 VC 维排序使函数嵌套的结构风险最小化

3.6.3　线性支持向量机

我们仍考虑标识过的训练数据$\{\boldsymbol{x}_i,y_i\}$，$i=1,2,\cdots,l$，其中 $\boldsymbol{x}_i\in\mathbb{R}^n$，$y_i\in\{-1,1\}$。现在讨论最简单的情况：在可分离数据上训练得到的线性学习机。

定义 3.6.6　设$\mathbb{R}^n$ 中有一个超平面 $\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}+b=0$，其中 $\boldsymbol{w}\in\mathbb{R}^n$ 是参数向量即超平面

的法线，$b\in\mathbb{R}$ 是截距，$|b|/\|\boldsymbol{w}\|$ 是超平面到原点的垂直距离(此处 $\|\cdot\|$ 是欧氏范数)，如果这个超平面可以把标识为正的样本集与标识为负的样本集分离开来，则称其为**分离超平面**(**separating hyperplane**)。设 d_+ 是分离超平面到正样本集的最短距离，而 d_- 是其到负样本集的最短距离。这个分离超平面的"**余度**"(**margin**)定义为 d_++d_-。

定义 3.6.7 所谓**线性支持向量机**(**linear support vector machine**)就是一种算法，以求得具有最大余度的分离超平面，即要求所有训练数据满足如下约束条件

$$\begin{cases}\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{w}+b\geqslant+1, & y_i=+1,\\ \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{w}+b\leqslant-1, & y_i=-1,\end{cases}\quad i=1,2,\cdots,l \tag{3-6-18}$$

或者写成一个式子

$$y_i(\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{w}+b)-1\geqslant 0,\quad \forall i \tag{3-6-19}$$

而以等式形式满足式(3-6-19)的点称为**支持向量**(**support vector**)。

现在我们考察式(3-6-18)，满足 $\boldsymbol{x}^{\mathrm{T}}\boldsymbol{w}+b=+1$ 的超平面是 H_1，其法线仍然是 $\boldsymbol{w}$，它到原点的垂直距离是 $|1-b|/\|\boldsymbol{w}\|$；而满足 $\boldsymbol{x}^{\mathrm{T}}\boldsymbol{w}+b=-1$ 的超平面是 H_2，其法线也是 $\boldsymbol{w}$，它到原点的垂直距离是 $|-1-b|/\|\boldsymbol{w}\|$；这是两个平行的超平面。而且 $d_+=d_-=1/\|\boldsymbol{w}\|$，余度是 $2/\|\boldsymbol{w}\|$。注意，在平行的两个超平面 H_1 和 H_2 之间没有任何训练点。

于是，线性支持向量机求取一对具有最大余度超平面的问题，即在满足式(3-6-19)的前提下对 $\|\boldsymbol{w}\|^2$ 求最小的问题；就是如下约束优化问题

$$\begin{cases}\min\limits_{\boldsymbol{w}}\dfrac{1}{2}\|\boldsymbol{w}\|^2\\ \text{s.t.}\quad y_i(\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{w}+b)-1\geqslant 0,\quad \forall i\end{cases} \tag{3-6-20}$$

这是一个典型的凸二次规划问题，因为目标函数本身是凸函数，而满足约束的点集也是凸集。

图 3-6-5 给出了二维情况下典型的线性分离直线。处在 H_1 或 H_2 上的点以等式形式满足式(3-6-19)，它们的变化影响余度的变化，即影响问题的解，这就是支持向量。

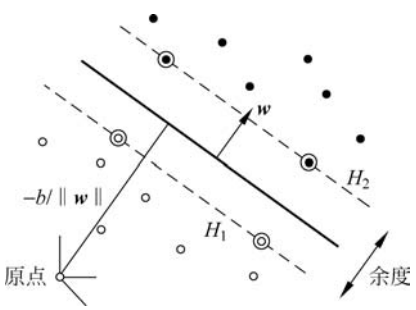


图 3-6-5 线性分离超平面：画圈的点是支持向量

现在把问题转换到 Lagrange 公式。这样做的必要性是：①上述优化问题中的约束条件能够用 Lagrange 乘子的约束来代替，而后者一般易于处理；②在把问题重新描述的过程中，训练数据将以向量点积的形式出现，便于推广到非线性情况。针对式(3-6-19)的不等式引入正 Lagrange 乘子 α_i，$i=1,2,\cdots,l$；于是给出 Lagrange 函数为

$$L_{\mathrm{P}}=\frac{1}{2}\|\boldsymbol{w}\|^2-\sum_{i=1}^{l}\alpha_i y_i(\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{w}+b)+\sum_{i=1}^{l}\alpha_i \tag{3-6-21}$$

而对于式(3-6-20)的优化问题，根据文献[66]中的定理 9.5.1，可以等价地求解其所谓 Wolfe 对偶问题，即

$$\begin{cases}\max\limits_{\boldsymbol{w},\boldsymbol{\alpha}} L_{\mathrm{P}}=\dfrac{1}{2}\|\boldsymbol{w}\|^2-\sum\limits_{i=1}^{l}\alpha_i y_i(\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{w}+b)+\sum\limits_{i=1}^{l}\alpha_i \\ \text{s.t.}\quad \dfrac{\partial L_{\mathrm{P}}}{\partial \boldsymbol{w}}=\boldsymbol{w}-\sum\limits_{i=1}^{l}\alpha_i y_i\boldsymbol{x}_i=0,\quad \alpha_i\geqslant 0,\quad i=1,2,\cdots,l\end{cases} \tag{3-6-22}$$

式中$\boldsymbol{\alpha}=(\alpha_1,\alpha_2,\cdots,\alpha_l)^{\mathrm{T}}$，其解与式(3-6-20)的解取得相同的 $\boldsymbol{w}$ 值。

注意，上式的约束条件等于给定方程

$$\boldsymbol{w}=\sum_{j=1}^{l}\alpha_j y_j\boldsymbol{x}_j \tag{3-6-23}$$

同时考虑到

$$\frac{\partial L_{\mathrm{P}}}{\partial b}=\sum_{i=1}^{l}\alpha_i y_i=0 \tag{3-6-24}$$

把以上两个式子代入式(3-6-21)，得

$$L_{\mathrm{D}}=\sum_{i=1}^{l}\alpha_i-\frac{1}{2}\sum_{i,j=1}^{l}\alpha_i\alpha_j y_i y_j\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{x}_j \tag{3-6-25}$$

于是，对于可分离的线性支持向量机的训练问题就是如下规划问题

$$\begin{cases}\max\limits_{\alpha} L_{\mathrm{D}}=\sum\limits_{i=1}^{l}\alpha_i-\dfrac{1}{2}\sum\limits_{i,j=1}^{l}\alpha_i\alpha_j y_i y_j\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{x}_j \\ \text{s.t.}\quad \sum\limits_{i=1}^{l}\alpha_i y_i=0,\quad \alpha_i\geqslant 0,\quad i=1,2,\cdots,l\end{cases} \tag{3-6-26}$$

把求解得到的最优$\boldsymbol{\alpha}$值代入式(3-6-23)就得到最优的 $\boldsymbol{w}$ 值。

对于约束优化问题，Karush-Kuhn-Tucker(KKT)条件无论从理论上或实践上都是非常重要的。对于上述约束优化问题，等价的 KKT 条件可以写成[67]

$$\begin{cases}\dfrac{\partial}{\partial w_j}L_{\mathrm{P}}=w_j-\sum\limits_{i=1}^{l}\alpha_i y_i x_{ij}=0, & j=1,2,\cdots,n \\ \dfrac{\partial}{\partial b}L_{\mathrm{P}}=-\sum\limits_{i=1}^{l}\alpha_i y_i=0 & \\ y_i(\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{w}+b)-1\geqslant 0, & i=1,2,\cdots,l \\ \alpha_i\geqslant 0,\quad \alpha_i(y_i(\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{w}+b)-1)=0, & \forall i\end{cases} \tag{3-6-27}$$

由此不仅可以得到最优的 $\boldsymbol{w}$ 值，而且可以求得最优的 α 值和 b 值。

一旦我们利用 l 个样本完成了对 SVM 的训练，我们如何来利用它对数据进行分类呢？实际上我们只是简单地确定测试数据 $\boldsymbol{x}$ 处在决策分界面(分类超平面)的哪一边。所以分类函数是

$$f(\boldsymbol{x})=\mathrm{sgn}(\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}+b) \tag{3-6-28}$$

其中 sgn(·)是符号函数；只是根据括号内的符号来确定 $\boldsymbol{x}$ 的模式。

以上讨论都是针对完全线性可分情况的。然而当数据非完全线性可分时，支持向量学习方法要通过构造一个**“软余度”**(**soft margin**)的分类超平面来达到最优分类的效果，如图 3-6-6 所示。

在非完全线性可分情况下，需要考虑分类误差带来的损失，这时在式(3-6-19)中引入一个松弛变量 $\xi_i\geqslant 0\quad i=1,\cdots,l$，成为

$$y_i(\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{w}+b)-1+\xi_i\geqslant 0,\quad \forall i \tag{3-6-29}$$

这时构造的“软余度”分类超平面是由如下优化问题确定

$$\begin{cases}\min\limits_{\boldsymbol{w}} & \dfrac{1}{2}\|\boldsymbol{w}\|^2+C\Big(\sum\limits_{i=1}^{l}\xi_i\Big)^k\\ \text{s.t.} & y_i(\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{w}+b)-1+\xi_i\geqslant 0,\quad i=1,\cdots,l\end{cases} \tag{3-6-30}$$

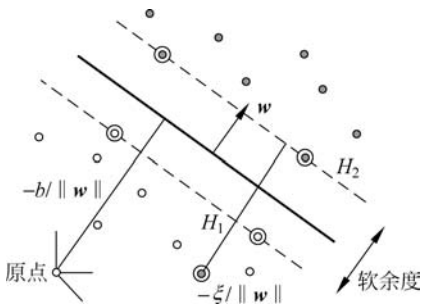


图 3-6-6 非完全线性可分时的分离超平面

式中 C 是对分类错误的惩罚因子，由使用者选择，用于调整置信范围和经验误差之间的均衡，较大的 C 意味着较小的经验误差，而小的 C 意味着更大的分类余度，而在数据不完全可分的情况下，参数 C 确定了经验风险的水平。而参数 k 也是可以选择的正整数，以保证是一个凸规划问题；$k=2$ 是二次规划问题。而 $k=1$ 具有更大的优越性，此时 Wolfe 对偶问题变为

$$\begin{cases}\max\limits_{\alpha} L_{\mathrm{D}}=\sum\limits_{i=1}^{l}\alpha_i-\dfrac{1}{2}\sum\limits_{i,j=1}^{l}\alpha_i\alpha_j y_i y_j\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{x}_j\\ \text{s.t.}\quad \sum\limits_{i=1}^{l}\alpha_i y_i=0,\quad 0\leqslant\alpha_i\leqslant C,\quad i=1,2,\cdots,l\end{cases} \tag{3-6-31}$$

问题的解则由下式给出

$$\boldsymbol{w}=\sum_{i=1}^{N_s}\alpha_i y_i\boldsymbol{x}_i \tag{3-6-32}$$

其中 N_s 就是支持向量数。这与最优超平面的差别仅仅是 α_i 有上界 C，如图 3-6-6 所示。

为了得到 KKT 条件，原问题的 Lagrange 函数变为

$$L_{\mathrm{P}}=\frac{1}{2}\|\boldsymbol{w}\|^2+C\sum_{i=1}^{l}\xi_i-\sum_{i=1}^{l}\alpha_i[y_i(\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{w}+b)-1+\xi_i]-\sum_{i=1}^{l}\mu_i\xi_i \tag{3-6-33}$$

其中 μ_i 是对于松弛变量 ξ_i 非负性的 Lagrange 乘子。于是，有原约束优化问题的 KKT 条件是

$$\begin{cases}\dfrac{\partial}{\partial w_j}L_{\mathrm{P}}=w_j-\sum\limits_{i=1}^{l}\alpha_i y_i x_{ij}=0,\quad j=1,2,\cdots,n\\ \dfrac{\partial}{\partial b}L_{\mathrm{P}}=-\sum\limits_{i=1}^{l}\alpha_i y_i=0\\ y_i(\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{w}+b)-1+\xi_i\geqslant 0,\qquad i=1,2,\cdots,l\\ \xi_i\geqslant 0,\quad \alpha_i\geqslant 0,\quad \mu_i\geqslant 0,\qquad i=1,2,\cdots,l\\ \alpha_i(y_i(\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{w}+b)-1+\xi_i)=0,\quad \forall i\\ \mu_i\xi_i=0,\qquad \forall i\end{cases} \tag{3-6-34}$$

类似地，我们可以利用最后两个方程来确定 b 的值。

3.6.4 非线性支持向量机

前面讨论的最优分类面都是线性情况，而实际中的大部分识别问题都不必是线性可

分的，其数据的分类需要非线性函数。如何把线性分类的方法推广应用到非线性情况，是SVM研究中关键的问题。此时要获得好的分类效果，必须采用非线性决策函数，统计学习理论采用如下的方法。

定义 3.6.8 所谓**非线性支持向量机**（**nonlinear support vector machine**）通过某种预先选择的非线性映射

$$\Phi:\mathcal{L}\to\mathcal{H} \tag{3-6-35}$$

进行变换，其中$\mathcal{L}=\mathbb{R}^n$是一个低维的欧氏空间，而$\mathcal{H}$是一个高维内积线性特征空间，一般是Hilbert空间；定义一个**核函数**（**kernel function**）K，使得

$$K(\boldsymbol{x}_i,\boldsymbol{x}_j)=\langle\Phi(\boldsymbol{x}_i),\Phi(\boldsymbol{x}_j)\rangle,\quad \forall\boldsymbol{x}_i,\boldsymbol{x}_j\in\mathcal{L} \tag{3-6-36}$$

其中$\langle\cdot,\cdot\rangle$表示$\mathcal{H}$中的内积，使得式(3-6-26)中的目标函数变为

$$L_{\mathrm{D}}=\sum_{i=1}^{l}\alpha_i-\frac{1}{2}\sum_{i,j=1}^{l}\alpha_i\alpha_j y_i y_j K(\boldsymbol{x}_i,\boldsymbol{x}_j) \tag{3-6-37}$$

这样就把低维空间的非线性分类问题转化为高维空间的线性分类问题，采用的方法与线性支持向量机相同。

注 在最优分类面中采用适当的核函数$K(\boldsymbol{x}_i,\boldsymbol{x}_j)$就可以实现某一非线性变换后的线性分类，而计算复杂度却没有增加。用式(3-6-37)代替式(3-6-26)中的目标函数，优化问题是完全相同的，仅仅把问题由低维空间的线性分类变成高维空间的线性分类。问题在于求得最优解需要得到高维空间$\mathcal{H}$的$\boldsymbol{w}$，即把式(3-6-23)的$\boldsymbol{x}_i$用$\Phi(\boldsymbol{x}_i)$代替，例题3.6.5将说明如何构造Φ。

例题 3.6.4 常用的一个核函数就是Gauss径向基函数（此时无须知道Φ的具体表示）

$$K(\boldsymbol{x}_i,\boldsymbol{x}_j)=\exp\{-\|\boldsymbol{x}_i-\boldsymbol{x}_j\|^2/(2\sigma^2)\}$$

在此特例中$\mathcal{H}$是一个无限维空间，此时直接应用Φ并不方便，在训练算法中可以处处用$K(\boldsymbol{x}_i,\boldsymbol{x}_j)$来代替$\Phi(\boldsymbol{x}_i)$与$\Phi(\boldsymbol{x}_j)$的内积，从而产生一个新的支持向量机对无限维空间进行分类。□

下面我们将讨论什么样的函数K是允许的。

定理 3.6.2（**Mercer条件**） 对于任意的对称函数$K(\boldsymbol{x},\boldsymbol{y})$，$\boldsymbol{x},\boldsymbol{y}\in\mathcal{L}$，以及一个映射$\Phi:\mathcal{L}\to\mathcal{H}$，它可以表示为特征空间$\mathcal{H}$中的内积运算，即$K(\boldsymbol{x},\boldsymbol{y})=\langle\Phi(\boldsymbol{x}),\Phi(\boldsymbol{y})\rangle$的充分必要条件是，对于任意不恒等于零的$g\in L^2(\mathcal{L})$，有下式成立

$$\int K(\boldsymbol{x},\boldsymbol{y})g(\boldsymbol{x})g(\boldsymbol{y})\mathrm{d}\boldsymbol{x}\mathrm{d}\boldsymbol{y}\geqslant 0 \tag{3-6-38}$$

证明 见文献[72]。■

例题 3.6.5 考虑$\mathcal{L}=\mathbb{R}^2$中的数据，我们选择$K(\boldsymbol{x}_i,\boldsymbol{x}_j)=(\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{x}_j)^2$，则容易找到空间$\mathcal{H}$，以及映射$\Phi:\mathbb{R}^2\to\mathcal{H}$，使得$(\boldsymbol{x}^{\mathrm{T}}\boldsymbol{y})^2=\langle\Phi(\boldsymbol{x}),\Phi(\boldsymbol{y})\rangle$。我们选择$\mathcal{H}=\mathbb{R}^3$，则有

$$\Phi(\boldsymbol{x})=[x_1^2,\sqrt{2}x_1x_2,x_2^2]^{\mathrm{T}}$$

我们定义$B=[0,1]\times[0,1]\subset\mathbb{R}^2$，它在$\Phi$的作用下$\mathcal{H}$中的象如图3-6-7所示。

对于B中的数据，经变换后到了高维空间。

注意，对于给定的核函数，Φ和$\mathcal{H}$都不是唯一的。

例如，对于$\mathcal{H}=\mathbb{R}^3$，可以选择 $\Phi(\boldsymbol{x})=\frac{1}{\sqrt{2}}[(x_1^2-x_2^2),2x_1x_2,(x_1^2+x_2^2)]^{\mathrm{T}}$；而且对于$\mathcal{H}=\mathbb{R}^4$，可以选择 $\Phi(\boldsymbol{x})=[x_1^2,x_1x_2,x_1x_2,x_2^2]^{\mathrm{T}}$。 □

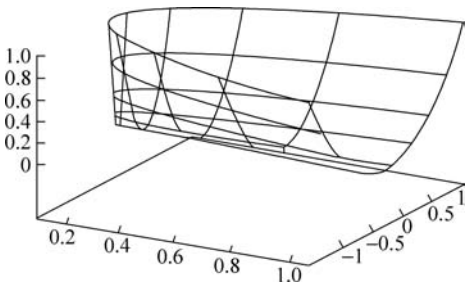


图 3-6-7　映射 $\boldsymbol{\Phi}$ 作用下 $\boldsymbol{B}$ 在$\mathcal{H}$中的象

在 SVM 完成训练后，相应于式(3-6-28)的分类函数为

$$f(\boldsymbol{x})=\mathrm{sgn}\left[\sum_{i=1}^{N_s}\alpha_i y_i K(\boldsymbol{s}_i,\boldsymbol{x})+b\right] \tag{3-6-39}$$

其中 $\boldsymbol{s}_i$ 是支持向量，sgn(·)是符号函数。上面描述的就是一般非线性情况下的支持向量机。

前面我们已经多次提到对映射 Φ 隐含处理的思想，但是我们也可以先构造 Φ，然后再得到核函数。下面的例题就是由 Φ 构造核函数的例子。

例题 3.6.6　考虑$\mathcal{L}=\mathbb{R}$，对于 $x\in\mathbb{R}$，设其可以进行 Fourier 展开并作前 N 项近似

$$g(x)=\frac{a_0}{2}+\sum_{i=1}^{N}[a_{1i}\cos(ix)+a_{2i}\sin(ix)]$$

而且能够视为$\mathcal{H}=\mathbb{R}^{2N+1}$ 中的内积，即设

$$\begin{cases}\boldsymbol{a}=[a_0/\sqrt{2},a_{11},\cdots,a_{1N},a_{21},\cdots,a_{2N}]^{\mathrm{T}}\in\mathbb{R}^{2N+1}\\ \Phi(x)=[1/\sqrt{2},\cos(x),\cdots,\cos(Nx),\sin(x),\cdots,\sin(Nx)]^{\mathrm{T}}\in\mathbb{R}^{2N+1}\end{cases}$$

则 $g(x)=\langle\boldsymbol{a},\Phi(x)\rangle=\boldsymbol{a}^{\mathrm{T}}\Phi(x)$；相应地，核函数可以写成

$$\langle\Phi(x_i),\Phi(x_j)\rangle=K(x_i,x_j)=\frac{\sin[(N+1/2)(x_i-x_j)]}{2\sin[(x_i-x_j)/2]}$$

这很容易证明。令 $\delta=x_i-x_j$，则有

$$\begin{aligned}\langle\Phi(x_i),\Phi(x_j)\rangle&=\frac{1}{2}+\sum_{l=1}^{N}\cos(lx_i)\cos(lx_j)+\sin(lx_i)\sin(lx_j)\\&=-\frac{1}{2}+\sum_{l=0}^{N}\cos(l\delta)=-\frac{1}{2}+\mathrm{Re}\left\{\sum_{l=0}^{N}\mathrm{e}^{\mathrm{j}l\delta}\right\}\\&=-\frac{1}{2}+\mathrm{Re}\{(1-\mathrm{e}^{\mathrm{j}(N+1)\delta})/(1-\mathrm{e}^{\mathrm{j}\delta})\}\\&=[\cos N\delta-\cos(N+1)\delta]/[2(1-\cos\delta)]\\&=[\sin((N+1/2)\delta)]/[2\sin(\delta/2)]\end{aligned}$$

这个方法对于很多用点积表示的数据都是适用的，只是表达形式有差异。 □

非线性 SVM 的常用核函数还有

$$K(\boldsymbol{x},\boldsymbol{y})=(\boldsymbol{x}^{\mathrm{T}}\boldsymbol{y}+1)^p \tag{3-6-40}$$

这是一个次方为 p 的多项式数据分类器，图 3-6-8 是一个 3 次分类器产生的结果。而

$$K(\boldsymbol{x},\boldsymbol{y})=\mathrm{e}^{-\|\boldsymbol{x}-\boldsymbol{y}\|^2/(2\sigma^2)} \tag{3-6-41}$$

是一个 Gauss 径向基函数(RBF)分类器，而且在此情况下通过 SVM 训练可以自动得到支持

向量 $\boldsymbol{s}_i$、权值向量$\boldsymbol{\alpha}$ 和截距 b 都能自动产生，而且可以获得比经典 RBF 更好的结果。同时

$$K(\boldsymbol{x},\boldsymbol{y})=\tanh(\kappa\boldsymbol{x}^{\mathrm{T}}\boldsymbol{y}-\delta) \tag{3-6-42}$$

是一种特殊的两层 sigmoid 神经网络分类器。

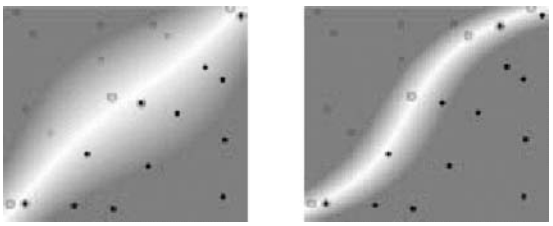


图 3-6-8　3 次多项式核函数产生的分界面：背景色表示决策面的形状

3.6.5　用于孤立点发现的 One-class SVM 算法

以上讨论的均是有监督的机器学习算法，能够处理的都是有标定的数据，1999 年 Bernhard Scholkopf 等提出一种能够处理无标定的数据的 One-class SVM 算法[72]，它是一种无监督学习的方法，主要用于异常检测和孤立点发现。孤立点发现的问题可描述如下。

定义 3.6.9　给定未标定训练数据 $\boldsymbol{x}_1,\cdots,\boldsymbol{x}_l\in\mathcal{L}$，其中 $l\in\mathbb{N}$ 是训练数据的样本数，$\mathcal{L}$是低维样本空间；假定训练数据中大部分具备某种特性，而很小部分属于孤立点，**One-class SVM** 就是要找到一个函数 $f(\boldsymbol{x})$，在大部分样本上取值为＋1，而在孤立点上取值为－1。其思路是通过核函数的选择，将低维样本空间变换到高维特征空间，然后在特征空间中寻找一个最优分类面，任一点的 $f(\boldsymbol{x})$值由其落在分类面的两侧来决定。同分类的 SVM 类似，这个问题的解由以下优化问题来得到

$$\begin{cases}\min\limits_{\boldsymbol{w},\xi_i,b}\dfrac{1}{2}\|\boldsymbol{w}\|^2+\dfrac{1}{vl}\sum\limits_{i=1}^{l}\xi_i-b\\ \text{s.t.}\quad \boldsymbol{w}^{\mathrm{T}}\Phi(\boldsymbol{x}_i)\geqslant b-\xi_i,\quad \xi_i\geqslant 0,\quad \forall i\end{cases} \tag{3-6-43}$$

其中 $v\in(0,1)$是个控制参数，Φ：$\mathcal{L}\to\mathcal{H}$是变换。

如果$(\boldsymbol{w}^{\mathrm{T}},b)^{\mathrm{T}}$ 是上面这个优化问题的解，则决策函数为

$$f(\boldsymbol{x})=\mathrm{sgn}(\boldsymbol{w}^{\mathrm{T}}\Phi(\boldsymbol{x})-b) \tag{3-6-44}$$

$f(\boldsymbol{x})$对数据集中大多数点取值为正，同时要求 $\|\boldsymbol{w}\|$ 比较小，而优化问题(3-6-43)中的参数 v 用于控制二者的折中。同样，对于问题(3-6-43)的解首先用核函数将原问题映射到特征空间，并采用 Lagrange 优化方法，得到原问题的对偶问题

$$\begin{cases}\min\limits_{\alpha}\quad \dfrac{1}{2}\sum\limits_{i,j=1}^{l}\alpha_i\alpha_j K(\boldsymbol{x}_i,\boldsymbol{x}_j)\\ \text{s.t.}\quad 0\leqslant\alpha_i\leqslant\dfrac{1}{vl},\quad \sum\limits_{i=1}^{l}\alpha_i=1\end{cases} \tag{3-6-45}$$

其中 $K(\boldsymbol{x}_i,\boldsymbol{x}_j)$是核函数，$\boldsymbol{\alpha}=(\alpha_1,\cdots,\alpha_l)^{\mathrm{T}}$。其最终判决函数 $f(\boldsymbol{x})$为

$$f(\boldsymbol{x})=\mathrm{sgn}\Big[\sum_{i=1}^{l}a_iK(\boldsymbol{x}_i,\boldsymbol{x})-b\Big] \tag{3-6-46}$$

这样，就通过核函数将输入空间转化到特征空间中，然后在特征空间中构造超平面，根据数据到原点的距离来实现分类，发现孤立点。

采用不同的核函数 $K(\boldsymbol{x},\boldsymbol{y})$，可以实现不同类型非线性决策面的学习机。核函数的形式主要有多项式核函数如式(3-6-40)、Gauss 核函数如式(3-6-41)等。

3.6.6 最小二乘支持向量机

J. A. K. Suykens 在 2001 年提出最小二乘支持向量机[84]，采用最小二乘线性系统作为损失函数，代替传统所采用的二次规划方法，取得了较好的效果。该方法运算简单，收敛速度快，精度高。算法描述如下。

定义 3.6.10 设训练样本集 $D=\{(\boldsymbol{x}_i,y_i):i=1,2,\cdots,l\}$，$\boldsymbol{x}_i\in\mathbb{R}^n$ 是输入数据，$y_i\in\mathbb{R}$ 是输出数据，**最小二乘支持向量机(least squares support vector machines，LS SVM)**描述为

$$\begin{cases}\min\limits_{\boldsymbol{w},b,e}J(\boldsymbol{w},\boldsymbol{e})=\dfrac{1}{2}\|\boldsymbol{w}\|^2+\dfrac{1}{2}\gamma\sum\limits_{i=1}^{l}e_i^2\\ \text{s.t.}\quad y_i=\boldsymbol{w}^{\mathrm{T}}\Phi(\boldsymbol{x}_i)+b+e_i,\quad i=1,2,\cdots,l\end{cases}\tag{3-6-47}$$

其中，$\Phi:\mathcal{L}=\mathbb{R}^n\to\mathcal{H}$ 是非线性映射；权值向量 $\boldsymbol{w}\in\mathcal{H}$，$b$ 是偏差量，$e_i\in\mathbb{R}$ 是误差，而 $\boldsymbol{e}=[e_1,\cdots,e_l]^{\mathrm{T}}\in\mathbb{R}^l$ 是误差向量，J 是损失函数，γ 是可调常数。

核空间映射函数的目的是从原始空间中抽取特征，将原始空间中的样本映射为高维特征空间中的一个向量，以解决原始空间中线性不可分的问题。

根据优化函数式(3-6-47)，定义 Lagrange 函数为

$$L(\boldsymbol{w},b,\boldsymbol{e},\boldsymbol{\alpha})=J(\boldsymbol{w},\boldsymbol{e})-\sum_{i=1}^{l}\alpha_i[\boldsymbol{w}^{\mathrm{T}}\Phi(\boldsymbol{x}_i)+b+e_i-y_i]\tag{3-6-48}$$

式中，$\alpha_i\in\mathbb{R}$ 是 Lagrange 乘子，$\boldsymbol{\alpha}=[\alpha_1,\cdots,\alpha_l]^{\mathrm{T}}\in\mathbb{R}^l$；$\boldsymbol{e}=[e_1,\cdots,e_l]^{\mathrm{T}}\in\mathbb{R}^l$。对上式进行优化，得

$$\begin{cases}\dfrac{\partial L}{\partial\boldsymbol{w}}=0\Rightarrow\boldsymbol{w}=\sum\limits_{i=1}^{l}\alpha_i\Phi(\boldsymbol{x}_i)\\ \dfrac{\partial L}{\partial b}=0\Rightarrow\sum\limits_{i=1}^{l}\alpha_i=0\\ \dfrac{\partial L}{\partial e_i}=0\Rightarrow\alpha_i=\gamma e_i,\quad i=1,2,\cdots,l\\ \dfrac{\partial L}{\partial\alpha_i}=0\Rightarrow\boldsymbol{w}^{\mathrm{T}}\Phi(\boldsymbol{x}_i)+b+e_i-y_i=0,\quad i=1,2,\cdots,l\end{cases}\tag{3-6-49}$$

消除变量 $\boldsymbol{w}$，$\boldsymbol{e}$，可得矩阵方程

$$\begin{bmatrix}0 & \mathbf{1}^{\mathrm{T}}\\ \mathbf{1} & \boldsymbol{\Omega}+\dfrac{1}{\gamma}\boldsymbol{I}\end{bmatrix}\begin{bmatrix}b\\ \boldsymbol{\alpha}\end{bmatrix}=\begin{bmatrix}0\\ \boldsymbol{y}\end{bmatrix}\tag{3-6-50}$$

式中，$\boldsymbol{y}=[y_1,\cdots,y_l]^{\mathrm{T}}$，$\mathbf{1}=[1,\cdots,1]^{\mathrm{T}}\in\mathbb{R}^l$，而且

$$\boldsymbol{\Omega}=\{\boldsymbol{\Omega}_{ij}\}_{l\times l},\quad \boldsymbol{\Omega}_{ij}=\Phi^{\mathrm{T}}(\boldsymbol{x}_j)\Phi(\boldsymbol{x}_i)=K(\boldsymbol{x}_j,\boldsymbol{x}_i) \tag{3-6-51}$$

于是利用最小二乘法可以得到参数$[b,\alpha^{\mathrm{T}}]^{\mathrm{T}}$的估计，其中核函数可以有不同的形式，如多项式核、多层感知(MLP)核、样条生成核和RBF核等。LS-SVM的预测函数为

$$y(\boldsymbol{x})=\sum_{i=1}^{l}\alpha_i K(\boldsymbol{x},\boldsymbol{x}_i)+b \tag{3-6-52}$$

文献[83]中采用最小二乘支持向量机对时间序列预测进行了研究，取得了良好的效果。

3.6.7 模糊支持向量机

在分类问题中我们往往关心的是对重要点的正确分类，并不关心一些像噪声之类的点是否被误分。这样，我们不再要求每个训练点精确属于两个类别中的某一个，而是以某种可能性属于某一类，这就引入模糊支持向量机的概念[75]。

定义 3.6.11 设给定一组数据集$D=\{(\boldsymbol{x}_i,y_i,s_i):i=1,2,\cdots,l\}$，其中$\boldsymbol{x}_i\in\mathbb{R}^n$是训练样本数据，$y_i\in\{-1,1\}$是标识数据，$s_i\in[\sigma,1]$是$\boldsymbol{x}_i$的隶属度，其中$\sigma>0$是无穷小量。定义$\Phi:\mathbb{R}^n\to\mathcal{Z}$是非线性映射，其中$\mathcal{Z}$是特征空间；**模糊支持向量机(fuzzy support vector machine，FSVM)**定义为如下优化问题

$$\begin{cases}\min & \dfrac{1}{2}\|\boldsymbol{w}\|^2+C\sum\limits_{i=1}^{l}s_i\xi_i\\ \text{s.t.} & y_i(\boldsymbol{w}^{\mathrm{T}}\boldsymbol{z}_i+b)\geqslant 1-\xi_i,\quad \xi_i\geqslant 0,\quad i=1,2,\cdots,l\end{cases} \tag{3-6-53}$$

其中$\boldsymbol{z}_i=\Phi(\boldsymbol{x}_i)$，$C$是一个常量，$\xi_i$是SVM中的误差度量，$s_i\xi_i$是不同权值的误差度量。

注意，一个很小的s_i减少了ξ_i的作用，因此，相应的x_i被认为是不重要的。为了求解这个最优问题，我们构造Lagrange函数

$$L(\boldsymbol{w},b,\boldsymbol{\xi},\boldsymbol{\alpha},\boldsymbol{\beta})=\frac{1}{2}\|\boldsymbol{w}\|^2+C\sum_{i=1}^{l}s_i\xi_i-\sum_{i=1}^{l}\alpha_i(y_i(\boldsymbol{w}^{\mathrm{T}}\boldsymbol{z}_i+b)-1+\xi_i)-\sum_{i=1}^{l}\beta_i\xi_i \tag{3-6-54}$$

这些参数必须满足下面的条件

$$\begin{cases}\dfrac{\partial L(\boldsymbol{w},b,\boldsymbol{\xi},\boldsymbol{\alpha},\boldsymbol{\beta})}{\partial \boldsymbol{w}}=\boldsymbol{w}-\sum\limits_{i=1}^{l}\alpha_i y_i\boldsymbol{z}_i=0\\ \dfrac{\partial L(\boldsymbol{w},b,\boldsymbol{\xi},\boldsymbol{\alpha},\boldsymbol{\beta})}{\partial b}=-\sum\limits_{i=1}^{l}\alpha_i y_i=0\\ \dfrac{\partial L(\boldsymbol{w},b,\boldsymbol{\xi},\boldsymbol{\alpha},\boldsymbol{\beta})}{\partial \xi_i}=s_iC-\alpha_i-\beta_i=0,\quad i=1,2,\cdots,l\end{cases} \tag{3-6-55}$$

将这些条件应用于Lagrange函数式(3-6-54)，则问题式(3-6-53)等价地变为

$$\begin{cases}\max J(\boldsymbol{\alpha})=\sum\limits_{i=1}^{l}\alpha_i-\dfrac{1}{2}\sum\limits_{i=1}^{l}\sum\limits_{j=1}^{l}\alpha_i\alpha_j y_i y_j K(\boldsymbol{x}_i,\boldsymbol{x}_j)\\ \text{s.t.}\quad \sum\limits_{i=1}^{l}y_i\alpha_i=0,\quad 0\leqslant\alpha_i\leqslant s_iC,\quad i=1,\cdots,l\end{cases} \tag{3-6-56}$$

类似地也可以得到Kuhn-Tucker条件。

3.6.8 小波支持向量机

定义 3.6.12 设支持向量机的核函数是由能够逼近任意非线性函数的多维小波函数构成，则称其为**小波核**(**wavelet kernal**)函数，而相应的支持向量机称为**小波支持向量机**(**wavelet support vector machine**，**WSVM**)[76]。设有小波母函数

$$h_{a,c}(x)=|a|^{-1/2}h\left(\frac{x-c}{a}\right) \tag{3-6-57}$$

此处 $x,a,c\in\mathbb{R}$，a 是伸缩因子，c 是转移因子。函数 $f\in L^2(\mathbb{R})$ 的小波变换可以写为

$$W_{a,c}(f)=\langle f(x),h_{a,c}(x)\rangle \tag{3-6-58}$$

其中$\langle\cdot,\cdot\rangle$表示 $L^2(\mathbb{R})$ 中的内积。式(3-6-58)表示函数 $f(x)$ 在小波基 $h_{a,c}(x)$ 上的分解。

对于任意小波母函数 $h(x)$，它必须满足下面的条件

$$W_h=\int_0^{\infty}\frac{|H(\omega)|^2}{|\omega|}\mathrm{d}\omega<\infty \tag{3-6-59}$$

其中 $H(\omega)$ 是 $h(x)$ 的 Fourier 变换。我们可以重构 $f(x)$ 如下

$$f(x)=\frac{1}{W_h}\int_{-\infty}^{\infty}\int_0^{\infty}W_{a,c}(f)h_{a,c}(x)\mathrm{d}a/a^2\mathrm{d}c \tag{3-6-60}$$

如果对上式取有限项近似，则有

$$\hat{f}(x)=\sum_{i=1}^{l}W_i h_{a_i,c_i}(x) \tag{3-6-61}$$

对于一个普通的多维小波函数，我们可以把它写成一维小波函数的乘积

$$h(\boldsymbol{x})=\prod_{i=1}^{N}h(x_i) \tag{3-6-62}$$

式中 $\boldsymbol{x}=(x_1,\cdots,x_n)^{\mathrm{T}}\in\mathbb{R}^n$。这里每个一维小波母函数必须满足式(3-6-59)。

定理 3.6.3 令 $h(x)$ 是一个小波母函数，a 和 c 分别表示伸缩和转移因子，$x,a,c\in\mathbb{R}$；如果 $\boldsymbol{x},\boldsymbol{x}'\in\mathbb{R}^n$，则内积小波核函数为

$$K(\boldsymbol{x},\boldsymbol{x}')=\prod_{i=1}^{n}h\left(\frac{x_i-c_i}{a}\right)h\left(\frac{x_i'-c_i'}{a}\right) \tag{3-6-63}$$

转移不变小波核函数(满足转移不变核法则)为

$$K(\boldsymbol{x},\boldsymbol{x}')=\prod_{i=1}^{n}h\left(\frac{x_i-x_i'}{a}\right) \tag{3-6-64}$$

证明 见文献[76]中附录 A。 ■

定理 3.6.4 考虑具有一般性的小波函数

$$h(x)=\cos(1.75x)(-x^2/2) \tag{3-6-65}$$

并给定伸缩因子 a，且 $a,x\in\mathbb{R}$；如果 $\boldsymbol{x},\boldsymbol{x}'\in\mathbb{R}^n$，则小波核函数为

$$K(\boldsymbol{x},\boldsymbol{x}')=\prod_{i=1}^{n}h\left(\frac{x_i-x_i'}{a}\right)=\prod_{i=1}^{n}\{\cos[1.75(x_i-x_i')/a]\exp[-\|x_i-x_i'\|/(2a^2)]\} \tag{3-6-66}$$

证明 见文献[76]中附录B。 ■

现在，我们给出WSVM作为分类器的决策函数

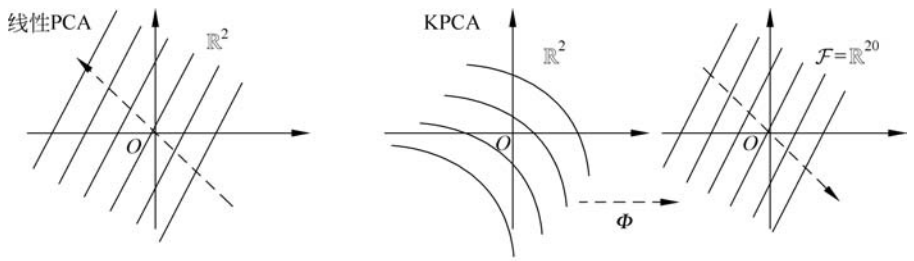
$$f(\boldsymbol{x})=\operatorname{sgn}\left\{\sum_{i=1}^{l}\alpha_i y_i\prod_{j=1}^{n}h[(x_j-x_j^{(i)})/a_i]+b\right\} \tag{3-6-67}$$

式中，$\boldsymbol{x}=(x_1,\cdots,x_n)^{\mathrm{T}}\in\mathbb{R}^n$，$x_j^{(i)}$ 表示第 i 个训练样本的第 j 个分量。

3.6.9 核主成分分析

定义 3.6.13 **核主成分分析（kernel principal component analysis，KPCA）**就是线性主成分分析的非线性泛化，它把原始数据空间非线性地映射到一个高维特征空间，并在这个特征空间中进行主成分分析。

图3-6-9给出了线性PCA与KPCA的比较图。



(a) 线性PCA　　(b) KPCA

图 3-6-9 线性 PCA 与 KPCA

设变换 Φ：$\mathbb{R}^n\to\mathcal{F}$ 实现了输入空间到特征空间的映射，即把输入空间的样本点 $\boldsymbol{x}_1$，$\boldsymbol{x}_2,\cdots,\boldsymbol{x}_l$ 变换为特征空间的样本点 $\Phi(\boldsymbol{x}_1),\Phi(\boldsymbol{x}_2),\cdots,\Phi(\boldsymbol{x}_l)$。假设[80]

$$\sum_{i=1}^{l}\Phi(\boldsymbol{x}_i)=0 \tag{3-6-68}$$

则$\mathcal{F}$空间上的协方差阵为

$$\boldsymbol{C}=\frac{1}{l}\sum_{j=1}^{l}\Phi(\boldsymbol{x}_j)\Phi^{\mathrm{T}}(\boldsymbol{x}_j) \tag{3-6-69}$$

现在求 $\boldsymbol{C}$ 的特征值 $\lambda\geqslant 0$ 和特征向量 $\boldsymbol{v}\in\mathcal{F}\backslash\{0\}$，

$$\lambda\boldsymbol{v}=\boldsymbol{C}\boldsymbol{v} \tag{3-6-70}$$

因为 $\boldsymbol{C}\boldsymbol{v}=\dfrac{1}{l}\sum_{j=1}^{l}\Phi(\boldsymbol{x}_j)<\Phi(\boldsymbol{x}_j),\boldsymbol{v}>=\lambda\boldsymbol{v}$，则所有非零特征值 $\lambda\neq 0$ 对应的特征向量 $\boldsymbol{v}$ 都在 $\Phi(\boldsymbol{x}_1),\Phi(\boldsymbol{x}_2),\cdots,\Phi(\boldsymbol{x}_l)$ 张成的平面内，从而存在不完全为零的系数 $\alpha_1,\alpha_2,\cdots,\alpha_l$，使得

$$\boldsymbol{v}=\sum_{i=1}^{l}\alpha_i\Phi(\boldsymbol{x}_i) \tag{3-6-71}$$

这样，对于 $j=1,2,\cdots,l$ 有

$$\lambda\langle\Phi(\boldsymbol{x}_j),\boldsymbol{v}\rangle=\langle\Phi(\boldsymbol{x}_j),\boldsymbol{C}\boldsymbol{v}\rangle=\lambda\sum_{i=1}^{l}\alpha_i\langle\Phi(\boldsymbol{x}_j),\Phi(\boldsymbol{x}_i)\rangle$$

$$=\frac{1}{l}\sum_{i=1}^{l}\alpha_i\langle\Phi(\boldsymbol{x}_j),\sum_{s=1}^{l}\Phi(\boldsymbol{x}_s)\rangle\langle\Phi(\boldsymbol{x}_s),\Phi(\boldsymbol{x}_i)\rangle \tag{3-6-72}$$

其中$\langle\cdot,\cdot\rangle$表示$\mathcal{F}$中的内积。

定义

$$\begin{cases}\boldsymbol{\alpha}\triangleq[\alpha_1,\alpha_2,\cdots,\alpha_l]^{\mathrm{T}}\\ \boldsymbol{K}\triangleq\{K_{ij}\}_{l\times l},\quad K_{ij}=\langle\Phi(\boldsymbol{x}_i),\Phi(\boldsymbol{x}_j)\rangle,\quad i,j=1,2,\cdots,l\end{cases} \tag{3-6-73}$$

则式(3-6-72)可以表示为

$$l\lambda\boldsymbol{K}\boldsymbol{\alpha}=\boldsymbol{K}\boldsymbol{K}\boldsymbol{\alpha} \tag{3-6-74}$$

显然,方程

$$l\lambda\boldsymbol{\alpha}=\boldsymbol{K}\boldsymbol{\alpha} \tag{3-6-75}$$

的解则必满足式(3-6-74)。通过对式(3-6-75)的求解,即可获得要求的特征值和特征向量。

图 3-6-10 给出了线性 KPCA 的一般原理。

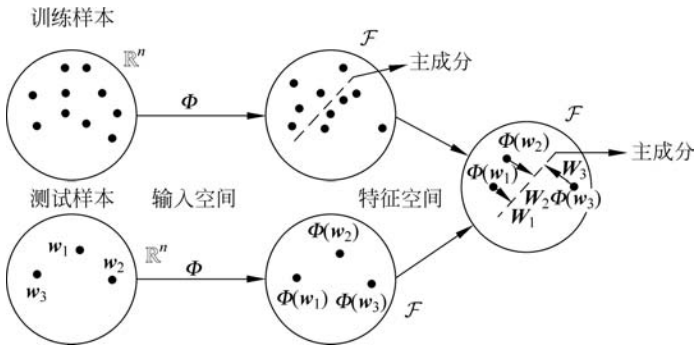


图 3-6-10 KPCA 的一般原理图

3.7 Bayes 网络基础

过去 30 多年间,Bayes 网络(BN)已经成为人工智能领域内不确定环境中进行知识表示和推理的一种有效工具[86-87]。BN 不仅对于大规模变量的联合概率分布提供了一种自然的紧凑的表示方式,而且也对于有效的概率推断提供了一个牢固的基础。虽然对于一类特殊的 BN 如树状网络或单连接网络存在多项式时间的推断算法,但对一般的网络却是一个 NP 难题。然而,在把 BN 应用于实际问题时,人们所面临的主要挑战就是如何设计一个简单而有效的近似推断算法,使得对于大规模的概率模型仍然能够满足实时性约束。目前已经发展了许多严格而又近似的 BN 推断算法,其中有些也是实时算法。本节将提供有关 BN 推断算法的基础知识,并不对具体算法进行深入讨论。

Bayes 网络也称为 **Bayes 置信网络(Bayesian belief network)**,或**因果网络(causal network)**,或**概率网络(probabilistic network)**,是当前人工智能领域内进行不确定性知识表示和推理的主要工具之一。

3.7.1 Bayes网络的一般概念

定理 3.7.1（**Bayes 定理**，Thomas Bayes，1763） Bayes 更新公式如下

$$p(H \mid E,c)=\frac{p(H \mid c)p(E \mid H,c)}{p(E \mid c)} \tag{3-7-1}$$

其中，H 是假设变量，E 是证据变量，c 是背景变量（可以不存在）；左边一项 $p(H|E,c)$ 称为后验概率，即考虑了 E 和 c 的影响之后 H 的概率；$p(H|c)$项称为只给定 c 时 H 的先验概率；$p(E|H,c)$是假定假设 H 和背景信息 c 为真的条件下证据 E 发生的概率，称为似然；而 $p(E|c)$是与假设 H 无关的证据 E 的概率。

证明 （略）。 ■

利用 Bayes 公式，我们可以把对假设的先验知识通过联合证据而更新成后验知识，从而达到更接近“真实”的目的，这就是 Bayes 推理的基础。

首先我们给出如下定义。

定义 3.7.1 一个 **Bayes 网络**（**Bayesian network，BN**）就是一个图，满足如下条件：

（1）一组随机变量$\{x_1,x_2,\cdots,x_n\}$构成了网络的结点，$V=\{1,2,\cdots,n\}$表示有限结点集合，而与之相应的随机变量构成随机向量 $\boldsymbol{x}=\{x_1,x_2,\cdots,x_n\}$；

（2）一组有向边（带箭头的连线）用于连接 V 中两两结点，由结点 k 指向结点 s 的箭头表示随机变量 x_k 直接影响随机变量 x_s；而 W 则表示各结点间有限个有向边的集合；

（3）每个结点都有一个**局部条件概率表**（**local conditional probability table，LCPT**），用于定量描述其父结点（指向它的结点）对该结点的作用，所有变量的 LCPT 构成**条件概率表**（**CPT**）；

（4）该图不存在有向环，因而称为**有向无环图**（**directed acyclic graph，DAG**）。

DAG 则定义了一个 Bayes 网络的结构。

为了理解 Bayes 网络的本质，我们先考虑对一个因果性问题进行建模。

例题 3.7.1 考虑下面一个所谓“判断妻子是否在家”问题，这是一个不确定判断问题。假定当某甲一个人晚上回家时，在进门以前希望知道妻子是否在家。根据经验，当他的妻子离开家时经常把前门的灯打开，但有时候她希望客人来时也打开这个灯。他们还养着一只狗，当无人在家时，这只狗被关在后院，而狗生病时也关在后院。如果狗在后院，就可以听见狗叫声，但有时听见的是邻居的狗叫。图 3-7-1 给出了这种逻辑关系的描述。

某甲可能利用这个图来预测他的妻子是否在家。如果妻子外出，狗也外出；如果灯是亮的，狗也不叫，妻子很可能外出。注意，这个例子的因果链并不是绝对的。也许，他的妻子外出时并没有带狗，也许没有开灯。有时我们能够利用这个图来推断，有时若证据不全时就很难推断。例如，如果灯是亮的但没有听到狗叫，是否能判定妻子在家呢？同样地，听到狗叫，但灯不亮又如何呢？一般情况下我们根据经验规定了相关事件发生的概率（见图 3-7-1 中的概率表，其中根结点规定的是先验概率，而非根结点规定的是条件概率；其中每个变量均取二值，如 a 表示该事件发生，$\bar{a}$ 表示该事件不发生）。利用这个网络，我们只能判断其妻子在家的可能性大小。

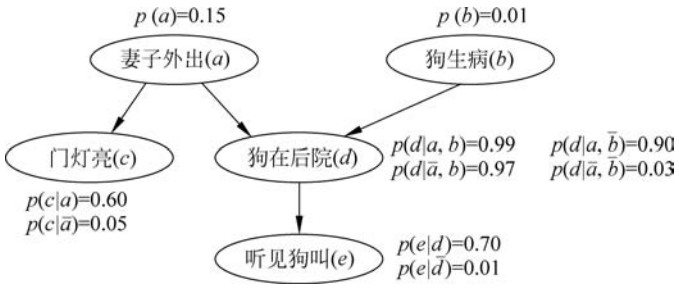


图 3-7-1 "判断妻子是否在家"问题的 Bayes 网络图 □

3.7.2 独立性假设

使用概率论的缺点之一就是荒谬地对大量的事件规定其完全概率分布，如果有 n 个随机变量，完全分布就要求有 2^n-1 个联合概率分布函数。这对规模较大的系统处理起来是不现实的。Bayes 网络必须建立独立性假设。

对于一个 Bayes 网络，随机变量 b 与其连接路径上最邻近的两个变量 a 和 c 之间存在三种连接方式，如图 3-7-2 所示。其中

① 直线连接(linear connection)：一个结点在其上，另一个结点在其下；

② 会聚连接(converging connection)：两个结点在其上；

③ 分叉连接(diverging connection)：两个结点在其下。

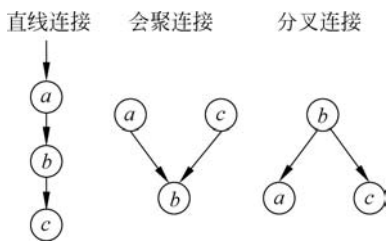


图 3-7-2 三种连接方式

以上相应于由 b 到 a 和 c 的箭头三种可能的组合。

下面给出所谓 d 连接路径的定义。

定义 3.7.2 在一个 Bayes 网络中，证据 $E=\{e_1,e_2,\cdots,e_m\}$ 定义为一组观测值。q 和 r 是网络中的两个结点，称由 q 到 r 的路径关于证据 E 是 **d 连接路径(d-connecting path)**，如果在该路径中的每个结点 s 具有如下特性之一：

(1) 是直线连接或分叉连接，而且 $s\notin E$；

(2) 是会聚连接，且 $s\in E$，或其后代结点属于 E。

定义 3.7.3 称变量 a 在给定证据 E(E 可以是空集，也可以非空但不包含 a 和 b)的条件下依赖于变量 b，如果给定 E 时，存在由 a 到 b 的一条 d 连接路径。若给定证据 E，变量 a 不依赖于(dependent on)变量 b，则称变量 a 在给定证据 E 时**独立于(independent on)**变量 b。

设 f 是任意随机变量，对于两个随机变量 a 和 b，给定 E 时彼此独立，但在给定 $E\cup\{f\}$ 有可能相互依赖。反之，在给定 E 时彼此依赖，但在给定 $E\cup\{f\}$ 有可能相互独立。如果两个随机变量 a 和 b 相互独立，简单地写成 $p(a|b)=p(a)$，而给定 e 时有可能使得 $p(a|b,e)\neq p(a|e)$。下面我们用例子来说明。

例题 3.7.2 考虑“判断妻子是否在家”的问题，给定“狗在后院”这一证据，由“妻子外出”到“听见狗叫”之间不存在 d 连接路径，实际上证据阻断了二者之间的连接。 □

在相关文献中，名词“d 分离”的应用更加普遍。

定义 3.7.4 两个结点称为是 **d 分离**（**d-separation**），如果它们之间不存在 d 连接路径。

粗略地说，两个结点之间是 d 连接的，要么它们之间就是因果路径（对应于条件 1），要么存在证据结点使得这两个结点之间相关（对应于条件 2）。

例题 3.7.3 为了理解这个定义，试想定义中的条件 1 不成立，那么我们就可以说，一个 d 连接路径可以不被证据所阻断。但这是不对的，因为已经看到，一旦知道了中间结点的状态，就不需要进一步知道任何事情。

参考图 3-7-1，我们说“狗生病”和“妻子外出”都能引起“狗在后院”这一状态的发生。但是，“狗生病”的概率依赖于“妻子外出”吗？根本不依赖。（我们可以想象存在某种依赖关系，但这可能要用其他 Bayes 网络来描述，不属于我们讨论的范围）。注意，它们二者只有一个路径，就是会聚到“狗在后院”。如果两个变量都能引起某个变量状态相同的变化，而它们二者之间又不存在任何联系，就说它们是相互独立的。于是，任何时候都有两种可能的原因引起同一状态变化的情形发生，这样我们就有了会聚结点。Bayes 网络常被用来决定是哪个因素引起状态的改变，会聚结点是常用的方法。

现在让我们来考虑定义中的条件 2。假定我们已经知道了“狗在后院”这一证据，此时“妻子外出”和“狗生病”是相互独立的吗？非也！虽然在没有证据时它们二者是相互独立的，但知道了“妻子在家”就增大了“狗生病”的概率，因为我们已经在很大程度上排除了“狗外出”的可能性，从而把“不太可能”变成“很大可能”。此时“妻子外出”和“狗生病”之间存在着 d 连接路径。这个路径通过“狗在后院”这个结点，而这个结点本身是一个条件结点。如果我们并没有“狗在后院”这一证据，而是仅仅“听见狗叫”。在此情况下，我们并不能肯定“狗在后院”，但是我们有了相关的证据“听见狗叫”，使得“狗在后院”的概率提高了。事实上，“听见狗叫”这个证据把它与会聚结点之上的两个结点联系起来了。条件 2 还说明，如果我们只有影响会聚结点的间接证据，也只能通过会聚结点形成路径。 □

如果一个结点随机变量取值的个数增多，CPT 的规模也将会随之增大，复杂的网络将会产生 NP 难题。然而在大多数情况下，实际的网络都由成百上千个结点组成，所以问题的简化将非常重要。

3.7.3 一致性概率

在建立 Bayes 网络时，一个常犯的错误就是规定的概率是非一致性的。

例题 3.7.4 考虑一个网络，其中有 $p(a|b)=0.7$，$p(b|a)=0.3$，且 $p(b)=0.5$，仔细看这些数据，好像没有出现什么差错。但是应用 Bayes 公式会有

$$p(a)p(b\mid a)/p(b)=p(a\mid b)\Rightarrow p(a)=p(b)p(a\mid b)/p(b\mid a)$$
$$\Rightarrow p(a)=0.5\times 0.7/0.3>1$$

这显然是错误的。 □

毋庸置疑，当网络的结点对应的概率取数很多时，问题就变得非常复杂，需要专门的技术来处理概率一致性问题。因此，对于 Bayes 网络的每个结点，给定与其父结点所有可能的组合数，必须要求：

(1) 这些数字符合一致性要求；

(2) 由 LCPT 规定的概率分布是唯一的。

下面会看到这个要求是正确的。首先要引入联合分布的概念。

定义 3.7.5 一组随机变量$\{x_1,x_2,\cdots,x_n\}$的**联合分布(joint distribution)**定义为

$$p(x_1,x_2,\cdots,x_n) \tag{3-7-2}$$

其所有可能的取值之和必须等于 1。

例题 3.7.5 考虑某一 Bayes 网络的两个随机变量$\{x_1,x_2\}$，而且它们都是二值变量；所有可能的联合分布是

$$p(a,b),\quad p(\bar{a},b),\quad p(a,\bar{b}),\quad p(\bar{a},\bar{b})$$

其中 $p(a,b)=p(x_1=a,x_2=b)$，而且 $\bar{a}$ 表示“非 a”；同时要求它们之和必须等于 1，即 $p(a,b)+p(\bar{a},b)+p(a,\bar{b})+p(\bar{a},\bar{b})=1$。从而有

$$p(a\mid b)=p(a,b)/p(b)=p(a,b)/[p(a,b)+p(\bar{a},b)]$$ □

一般而言，对于 n 个随机变量$\{x_1,x_2,\cdots,x_n\}$，假定它们都是二值变量，所有可能的联合分布有 2^n 个，而且要求

$$\sum_{2^n\text{可能值}} p(x_1,x_2,\cdots,x_n)=1 \tag{3-7-3}$$

对于一个 Bayes 网络而言，其联合概率分布由每个随机变量的分布的乘积唯一地进行定义。

设 $S\subseteq V$ 是 Bayes 网络部分结点构成的集合，$\boldsymbol{x}_s$ 表示与之相应的随机向量；Bayes 网络的**结构图(graph of structure)**就规定了这些变量之间的条件独立关系。设图形结构表示为 $G=(G_1,G_2,\cdots,G_n)$，其中 $G_i\subseteq V$ 表示结点 i 的父结点集合，在给定图形结构的前提下，变量 $\boldsymbol{x}=\{x_1,x_2,\cdots,x_n\}$的概率构成如下

$$p(\boldsymbol{x}\mid G,\theta)=\prod_{i=1}^{n} p(x_i\mid \boldsymbol{x}_{G_i},\theta) \tag{3-7-4}$$

其中 $\boldsymbol{x}_{G_i}$ 是 x_i 所有父结点随机变量构成的随机向量，θ 是相关的参数，而 $p(x_i\mid\boldsymbol{x}_{G_i},\theta)$ 是局部条件概率分布。

例题 3.7.6 考虑图 3-7-1 的 Bayes 网络，来检查其概率一致性

(1) 结点“a”：$p(a)=0.15,p(\bar{a})=0.85\Rightarrow p(a)+p(\bar{a})=1$

(2) 结点“b”：$p(b)=0.01,p(\bar{b})=0.09\Rightarrow p(b)+p(\bar{b})=1$

(3) 结点“c”：

$$\begin{aligned}
&p(c\mid a)=0.60,p(c\mid\bar{a})=0.05\Rightarrow p(\bar{c}\mid a)=0.40,p(\bar{c}\mid\bar{a})=0.95\\
&\qquad\Rightarrow p(a,c)=p(c\mid a)p(a)=0.60\times0.15=0.09,\\
&p(\bar{a},c)=p(c\mid\bar{a})p(\bar{a})=0.05\times0.85=0.0425,\\
&p(a,\bar{c})=p(\bar{c}\mid a)p(a)=0.40\times0.15=0.06,\\
&p(\bar{a},\bar{c})=p(\bar{c}\mid\bar{a})p(\bar{a})=0.95\times0.85=0.8075,
\end{aligned}$$

$$\Rightarrow p(a,c)+p(\bar{a},c)+p(a,\bar{c})+p(\bar{a},\bar{c})$$
$$=0.09+0.0425+0.06+0.8075=1$$

(4) 结点“d”：

$$p(d\mid a,b)=0.99,p(d\mid a,\bar{b})=0.90,p(d\mid\bar{a},b)=0.97,p(d\mid\bar{a},\bar{b})=0.03,$$
$$\Rightarrow p(\bar{d}\mid a,b)=0.01,p(\bar{d}\mid a,\bar{b})=0.10,p(\bar{d}\mid\bar{a},b)=0.03,p(\bar{d}\mid\bar{a},\bar{b})=0.97$$
$$\Rightarrow p(a,b,d)=p(d\mid a,b)p(a)p(b)=0.99\times0.15\times0.01=0.001\,485,$$
$$p(a,\bar{b},d)=p(d\mid a,\bar{b})p(a)p(\bar{b})=0.90\times0.15\times0.99=0.133\,65,$$
$$p(\bar{a},b,d)=p(d\mid\bar{a},b)p(\bar{a})p(b)=0.97\times0.85\times0.01=0.008\,245,$$
$$p(\bar{a},\bar{b},d)=p(d\mid\bar{a},\bar{b})p(\bar{a})p(\bar{b})=0.03\times0.85\times0.99=0.025\,245,$$
$$p(a,b,\bar{d})=p(\bar{d}\mid a,b)p(a)p(b)=0.01\times0.15\times0.01=0.000\,015,$$
$$p(a,\bar{b},\bar{d})=p(\bar{d}\mid a,\bar{b})p(a)p(\bar{b})=0.10\times0.15\times0.99=0.014\,85,$$
$$p(\bar{a},b,\bar{d})=p(\bar{d}\mid\bar{a},b)p(\bar{a})p(b)=0.03\times0.85\times0.01=0.000\,255,$$
$$p(\bar{a},\bar{b},\bar{d})=p(\bar{d}\mid\bar{a},\bar{b})p(\bar{a})p(\bar{b})=0.97\times0.85\times0.99=0.816\,255$$
$$\Rightarrow p(a,b,d)+p(a,\bar{b},d)+p(\bar{a},b,d)+p(\bar{a},\bar{b},d)+$$
$$p(a,b,\bar{d})+p(a,\bar{b},\bar{d})+p(\bar{a},b,\bar{d})+p(\bar{a},\bar{b},\bar{d})$$
$$=0.001\,485+0.133\,65+0.008\,245+0.025\,245+$$
$$0.000\,015+0.014\,85+0.000\,255+0.816\,255=1$$

(5) 结点“e”：

$$p(e\mid d)=0.70,p(e\mid\bar{d})=0.01,p(d)=0.168\,625,p(\bar{d})=0.831\,375$$
$$\Rightarrow p(\bar{e}\mid d)=0.30,p(\bar{e}\mid\bar{d})=0.99$$
$$\Rightarrow p(d,e)=p(e\mid d)p(d)=0.70\times0.168\,625=0.118\,037\,5$$
$$p(\bar{d},e)=p(e\mid\bar{d})p(\bar{d})=0.01\times0.831\,375=0.008\,313\,75$$
$$p(d,\bar{e})=p(\bar{e}\mid d)p(d)=0.30\times0.168\,625=0.050\,587\,5$$
$$p(\bar{d},\bar{e})=p(\bar{e}\mid\bar{d})p(\bar{d})=0.99\times0.831\,375=0.823\,061\,25$$
$$\Rightarrow p(d,e)+p(\bar{d},e)+p(d,\bar{e})+p(\bar{d},\bar{e})$$
$$=0.118\,037\,5+0.008\,313\,75+0.050\,587\,5+0.823\,061\,25=1$$

所以满足概率一致性要求。同时得到给定网络结构和条件概率表时变量的联合分布。

□

3.7.4 Bayes 网络推断

一个 Bayes 网络可以看作一个概率专家系统，其中概率知识基础由网络拓扑以及每个结点的 LCPT 表示。建立这个知识基础的主要目的是用于推断，即计算产生对该领域问题的解答。Bayes 网络主要有两种推断方式，即置信更新和置信修正。

定义 3.7.6 设 E 表示所有**证据结点**(**evidence node**)，λ 表示证据结点上的观测值，y 表示所有**询问结点**(**query node**)，所谓**置信更新**(**belief update**)，或称为**概率推断**

(**probability inference**),就是在给定 λ 的条件下,求 $\boldsymbol{y}$ 的后验概率分布

$$p(\boldsymbol{y} \mid \lambda) \tag{3-7-5}$$

可以由 CTP 得到 $\boldsymbol{y}$ 的先验分布 $p(\boldsymbol{y})$,以及似然 $p(\lambda \mid \boldsymbol{y})$,从而可以由 Bayes 公式得

$$p(\boldsymbol{y} \mid \lambda) = \frac{p(\lambda \mid \boldsymbol{y}) p(\boldsymbol{y})}{\sum_{\boldsymbol{y}} p(\lambda \mid \boldsymbol{y}) p(\boldsymbol{y})} \tag{3-7-6}$$

而对于 $\boldsymbol{y}$ 的任意函数 $f(\boldsymbol{y})$,也可以推断其后验期望,即

$$E[f(\boldsymbol{y}) \mid \lambda] = \frac{\sum_{\boldsymbol{y}} f(\boldsymbol{y}) p(\lambda \mid \boldsymbol{y}) p(\boldsymbol{y})}{\sum_{\boldsymbol{y}} p(\lambda \mid \boldsymbol{y}) p(\boldsymbol{y})} \tag{3-7-7}$$

定义 3.7.7 所谓**置信修正**(**belief revision**)就是在给定观测证据的前提下,对某些假设变量求取最有可能的例证;产生的结果就是假设变量的一个最优例证表。

很多置信更新算法经过很小的修改就能用于置信修正,反之亦然。本节只讨论置信更新算法。置信更新的计算,在使用 Bayes 网络时最大的困难就是在一般情况下这是一个 NP 难题。其次,随着网络规模的增大,计算量则按指数增长。解决这个问题的思路就是:要么系统规模很小,一般只有数十个结点;要么结点成千上万,而计算时间可以任意延长。

首先,我们必须澄清是否需要得到精确解。精确解在一般情况下是 NP 难题,近似解则要求与精确解非常接近,要以很高的概率在正确解的很小邻域内。

首先我们必须从最简单的问题开始研究精确解问题。

定义 3.7.8 所谓**单连接网络**(**singly connected network**)或称**多叉树网络**(**polytree network**),就是在相应的基础无向图上任意两个结点之间至多有一个连接路径。

所谓**基础无向图**(**underlying undirected graph**)就是忽略箭头后的网络图。

例题 3.7.7 考虑图 3-7-3 和图 3-7-4 的网络,其相应的基础无向图(不管其方向)任意两点间至多存在一条路径,所以是单连接网络。

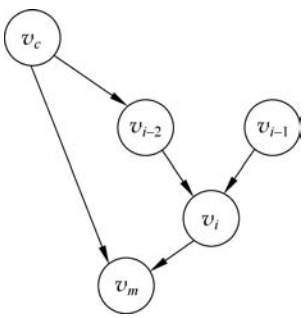


图 3-7-3 非单连接网络

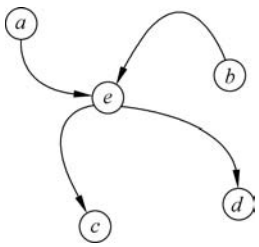


图 3-7-4 单连接网络的关系

而图 3-7-3 中,在结点 v_c 和 v_m 之间,除了路径 $v_c - v_m$ 之外,还有路径 $v_c - v_{i-2} - v_i - v_m$,所以是非单连接网络。 □

对于单连接网络,假定我们只考虑单个询问结点 y,我们针对图 3-7-5 的情况分别加以讨论。

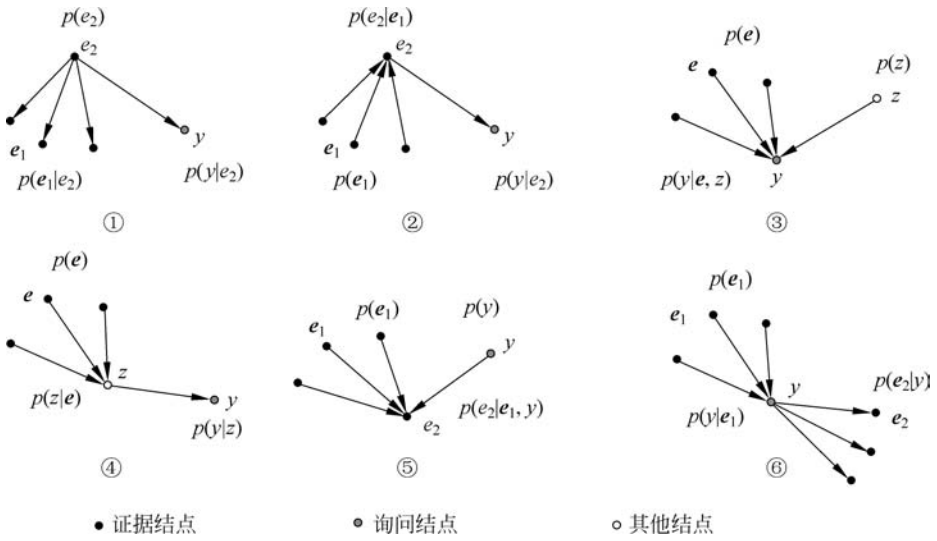


图 3-7-5　单连接网络进行推断的几种连接方式

图中，①、②、③均为 d 分离的，所以比较简单，即

① 分叉连接，但 $e_2 \in E$。此时如果有证据 $\boldsymbol{e}_1=\lambda_1, e_2=\lambda_2$，则

$$p(y \mid \lambda)=p(y \mid \lambda_1,\lambda_2)=p(y \mid \lambda_2) \tag{3-7-8}$$

② 直线连接，但 $e_2 \in E$。此时如果有证据 $\boldsymbol{e}_1=\lambda_1, e_2=\lambda_2$，则

$$p(y \mid \lambda)=p(y \mid \lambda_1,\lambda_2)=p(y \mid \lambda_2) \tag{3-7-9}$$

③ 会聚连接，但 $y \notin E$。此时如有证据 $\boldsymbol{e}=\lambda$，则

$$p(y \mid \lambda)=\sum_z p(y \mid \lambda, z) \tag{3-7-10}$$

而其中④、⑤、⑥ 均为 d 连接的，所以相对比较复杂，即

④ 直线连接，且 $z \notin E$。此时如果有证据 $\boldsymbol{e}=\lambda$，则

$$p(y \mid \lambda)=\frac{p(y,\lambda)}{p(\lambda)}=\frac{\sum_z p(y,z,\lambda)}{p(\lambda)}=\frac{\sum_z p(y \mid z)p(z \mid \lambda)p(\lambda)}{p(\lambda)}$$
$$=\sum_z p(y \mid z)p(z \mid \lambda) \tag{3-7-11}$$

⑤ 会聚连接，且 $e_2 \in E$。此时如果有证据 $\boldsymbol{e}_1=\lambda_1, e_2=\lambda_2$，则

$$p(y \mid \lambda)=p(y \mid \lambda_1,\lambda_2)=\frac{p(y,\lambda_1,\lambda_2)}{p(\lambda_1,\lambda_2)}=\frac{p(\lambda_2 \mid \lambda_1, y)p(\lambda_1)p(y)}{\sum_y p(\lambda_2 \mid \lambda_1, y)p(\lambda_1)p(y)}$$
$$=\frac{p(\lambda_2 \mid \lambda_1, y)p(y)}{\sum_y p(\lambda_2 \mid \lambda_1, y)p(y)} \tag{3-7-12}$$

⑥ 直线连接，且 $y \notin E$。这是比较复杂的一种情况。此时如果有证据 $\boldsymbol{e}_1=\lambda_1, \boldsymbol{e}_2=\lambda_2$，则有

$$p(y \mid \lambda)=p(y \mid \lambda_1,\lambda_2)=\frac{p(y \mid \lambda_1)p(\lambda_2 \mid y)}{\sum_y p(y \mid \lambda_1)p(\lambda_2 \mid y)} \tag{3-7-13}$$

其中右边分子上第一项为向上连接的后验概率，第二项为向下连接的后验概率，而分母则是分子对所有可能的 y 值求和。这将使得问题得以大大简化，现在证明如下：

首先根据 Bayes 规则，得

$$p(y \mid \lambda_1,\lambda_2)=\frac{p(y,\lambda_1,\lambda_2)}{p(\lambda_1,\lambda_2)}=\frac{p(y)p(\lambda_1 \mid y)p(\lambda_2 \mid \lambda_1,y)}{p(\lambda_1,\lambda_2)}$$

其次，注意

$$p(\lambda_2 \mid \lambda_1,y)=p(\lambda_2 \mid y)$$

这是因为，根据单连接条件，y 把 λ_1 和 λ_2 分开来了，从而得到

$$p(y \mid \lambda_1,\lambda_2)=\frac{p(y)p(\lambda_1 \mid y)p(\lambda_2 \mid y)}{p(\lambda_1)p(\lambda_2 \mid \lambda_1)}$$

重新调整各项，得到

$$p(y \mid \lambda_1,\lambda_2)=\frac{p(y)p(\lambda_1 \mid y)}{p(\lambda_1)}\cdot\frac{p(\lambda_2 \mid y)}{p(\lambda_2 \mid \lambda_1)}$$

再对上式右边第一个分式应用 Bayes 规则，可得

$$p(y \mid \lambda_1,\lambda_2)=\frac{p(y \mid \lambda_1)p(\lambda_2 \mid y)}{p(\lambda_2 \mid \lambda_1)}$$

再利用式(3-7-11)，即可得到式(3-7-13)。 □

例题 3.7.8 仍考虑图 3-7-1 的 Bayes 网络，假定我们已经知道“妻子外出”并“听见狗叫”，问“狗在后院”的后验概率是多少？

此时，$\lambda_1=a$，$\lambda_2=e$，$y=d$，

$$\begin{aligned}p(y=d \mid \lambda)&=p(d \mid a)=p(d \mid a,b)p(b)+p(d \mid a,\bar{b})p(\bar{b})\\&=0.99\times 0.01+0.9\times 0.99=0.900\,801\end{aligned}$$

$$p(\lambda_2 \mid y=d)=p(e \mid d)=0.7$$

而

$$\begin{aligned}p(y=\bar{d} \mid \lambda_1)&=p(\bar{d} \mid a)=p(\bar{d} \mid a,b)p(b)+p(\bar{d} \mid a,\bar{b})p(\bar{b})\\&=0.01\times 0.01+0.1\times 0.99=0.0991\end{aligned}$$

$$p(\lambda_2 \mid y=\bar{d})=p(e \mid \bar{d})=0.01$$

所以根据式(3-7-12)有

$$\begin{aligned}p(y \mid \lambda_1,\lambda_2)&=\frac{p(y \mid \lambda_1)p(\lambda_2 \mid y)}{\sum_y p(y \mid \lambda_1)p(\lambda_2 \mid y)}\\&=\frac{0.900\,801\times 0.7}{0.900\,801\times 0.7+0.0991\times 0.01}=0.998\,431\end{aligned}$$

即在知道“妻子外出”并“听见狗叫”的前提下，“狗在后院”的可能性是 0.998 431。 □

例题 3.7.9 再次考虑图 3-7-1 的 Bayes 网络，假定我们已经知道“妻子外出”并知道“狗在后院”，问“狗生病”的后验概率是多少？

此时，$\lambda_1=a$，$\lambda_2=d$，$y=b$，

$$p(\lambda_2 \mid \lambda_1,y=b)=p(d \mid a,b)=0.99,\quad p(y=b)=0.01$$

而

$$p(\lambda_2 \mid \lambda_1, y=\bar{b}) = p(d \mid a, \bar{b}) = 0.90, \quad p(y=\bar{b}) = 0.99$$

根据式(3-7-11)则有

$$p(y \mid \lambda_1, \lambda_2) = \frac{p(\lambda_2 \mid \lambda_1, y)p(y)}{\sum_y p(\lambda_2 \mid \lambda_1, y)p(y)} = \frac{0.99 \times 0.01}{0.99 \times 0.01 + 0.90 \times 0.99} = 0.011$$

即在知道“妻子外出”并知道“狗在后院”的前提下，“狗生病”的可能性是 0.011。

这个后验概率比较接近先验概率。但是，如果改变条件概率

$$p(\lambda_2 \mid \lambda_1, y=\bar{b}) = p(d \mid a, \bar{b}) = 0.09$$

即在“妻子外出”和“狗生病”同时发生时，“狗在后院”的可能性是 0.09，则在知道“妻子外出”并知道“狗在后院”的前提下，“狗生病”的可能性变为

$$p(y \mid \lambda_1, \lambda_2) = \frac{p(\lambda_2 \mid \lambda_1, y)p(y)}{\sum_y p(\lambda_2 \mid \lambda_1, y)p(y)} = \frac{0.99 \times 0.01}{0.99 \times 0.01 + 0.09 \times 0.99} = 0.10$$

使得可能性提高了一个数量级。 □

关于 Bayes 网络推断有许多精确方法和近似方法，而且有所谓参数自适应和结构自适应的方法，此处不再详细讨论。

3.8 大数据时代的云计算

正如第 1 章末所述，随着互联网技术和云计算技术等的飞速发展，信息领域的大数据时代已经来临。我们处理的多源信息融合技术也随着大数据时代的到来发生了重大变化。现代化的军事应用或工业应用系统，大都利用互联网和云计算技术来处理日益复杂的信息融合问题。本节讨论大数据时代的云计算及其对信息融合技术实现的影响。

3.8.1 云计算的概念

近年来，云计算(cloud computing)研究成为各大领域的关注热点。很多学者和专业人士逐步将云计算作为计算机网络技术应用架构的一个核心。云计算服务所提供的庞大网络数据中心不仅存储着大部分的应用软件和数据信息，还掌握着应用程序的管理及信息数据的安全维护工作。在为用户提供查询便利的同时，也可消除诸多安全隐患。

对云计算的定义有多种说法。现阶段广为接受的是美国国家标准与技术研究院(NIST)的定义：云计算是一种按使用量付费的模式，这种模式提供可用的、便捷的、按需的网络访问，进入可配置的计算资源共享池(资源包括网络、服务器、存储、应用软件和服务等)，这些资源能够被快速提供，只需投入很少的管理工作或与服务供应商进行很少的交互。

云计算通常通过互联网来提供动态易扩展且经常是虚拟化的资源。云是网络、互联网的一种比喻说法，云计算甚至可以实现每秒 10 万亿次的运算能力，拥有这么强大的计算能力可以模拟核爆炸、预测气候变化和市场发展趋势，当然可以用来快速实现网络化的多源信息融合。在许多情况下，用户可以通过计算机、手机等方式接入数据融合中心，

按自己的需求进行运算处理。

云计算常与网格计算、效用计算、自主计算相混淆。网格计算是分布式计算的一种，是由一群松散耦合的计算机组成的一个超级虚拟计算机，常用来执行一些大型任务；效用计算是IT资源的一种打包和计费方式，比如按照计算、存储分别计量费用，像传统的电力等公共设施一样；自主计算是具有自我管理功能的计算机系统。事实上，云计算部署依赖于计算机集群(但与网格的组成、体系结构、目的、工作方式大相径庭)，也吸收了自主计算和效用计算的特点。

换句话说，云计算是分布式计算(distributed computing)、并行计算(parallel computing)、效用计算(utility computing)、网络存储(network storage technologies)、虚拟化(virtualization)、负载均衡(load balance)、热备份冗余(high available)等传统计算机和网络技术发展融合的产物。云计算是继20世纪80年代大型计算机到客户端-服务器的大转变之后的又一种巨变。

被普遍接受的云计算特点如下：

(1) 超大规模。“云”具有相当的规模，Google云计算已经拥有100多万台服务器，Amazon、IBM、微软、雅虎等的“云”均拥有几十万台服务器。企业私有云一般拥有数百上千台服务器。“云”能赋予用户前所未有的计算能力。

(2) 虚拟化。云计算支持用户在任意位置使用各种终端获取应用服务。所请求的资源来自“云”，而不是固定的有形的实体。应用在“云”中某处运行，但实际上用户无须了解，也不用担心应用运行的具体位置。只需要一台笔记本或者一个手机，就可以通过网络服务来实现需要的一切，甚至包括超级计算这样的任务。

(3) 高可靠性。“云”使用了数据多副本容错、计算结点同构可互换等措施来保障服务的高可靠性，使用云计算比使用本地计算机可靠。

(4) 通用性。云计算不针对特定的应用，在“云”的支撑下可以构造出千变万化的应用，同一个“云”可以同时支撑不同的应用运行。

(5) 高可扩展性。“云”的规模可以动态伸缩，满足应用和用户规模增长的需要。

(6) 按需服务。“云”是一个庞大的资源池，用户按需购买；云可以像自来水、电、煤气那样计费。

(7) 极其廉价。由于“云”的特殊容错措施可以采用极其廉价的结点来构成云，“云”的自动化集中式管理使大量企业无须负担日益高昂的数据中心管理成本，“云”的通用性使资源的利用率较之传统系统大幅提升，因此用户可以充分享受“云”的低成本优势，经常只要花费很少的资金、几天时间就能完成以前需要大量资金、数月才能完成的任务。

(8) 潜在的危险性。云计算服务除了提供计算服务外，还必然提供了存储服务。但是云计算服务当前垄断在私人机构(企业)手中，仅仅能够提供商业信用。对于政府机构、商业用户、军事机构选择云计算服务应保持足够的警惕。一旦商业用户大规模使用私人机构提供的云计算服务，无论其技术优势有多强，都不可避免地让这些私人机构以“数据(信息)”的重要性挟制整个社会。对于信息社会而言，“信息”是至关重要的。另一方面，云计算中的数据对于数据所有者以外的其他云计算用户是保密的，但是对于提供云计算的商业机构而言确实毫无秘密可言。所有这些潜在的危险，是政府机构、商业机和军事机构选择云计算服务，特别是国外机构提供的云计算服务时，不得不考虑的一个

重要的前提。

3.8.2 云计算的快速发展

2006年8月9日，Google首席执行官埃里克·施密特(Eric Schmidt)在搜索引擎大会(SES San Jose 2006)上首次提出“云计算”的概念。2007年10月，Google与IBM开始在美国大学校园，包括卡内基梅隆大学、麻省理工学院、斯坦福大学、加州大学伯克利分校及马里兰大学等，推广云计算的计划，这项计划希望能降低分布式计算技术在学术研究方面的成本，并为这些大学提供相关的软硬件设备及技术支持。而学生则可以通过网络开发各项以大规模计算为基础的研究计划。2008年1月30日，Google宣布在台湾地区启动“云计算学术计划”，将与台湾大学、台湾交通大学等学校合作，将这种先进的大规模、快速将云计算技术推广到校园。2008年2月1日，IBM宣布将在中国无锡太湖新城科教产业园为中国的软件公司建立全球第一个云计算中心(Cloud Computing Center)。2008年7月29日，雅虎、惠普和英特尔宣布一项涵盖美国、德国和新加坡的联合研究计划，推出云计算研究测试床，推进云计算。该计划要与合作伙伴创建6个数据中心作为研究试验平台，每个数据中心配置1400～4000个处理器。这些合作伙伴包括新加坡资讯通信发展管理局、德国卡尔斯鲁厄大学Steinbuch计算中心、美国伊利诺伊大学香槟分校、英特尔研究院、惠普实验室和雅虎。2008年8月3日，美国专利商标局网站信息显示，戴尔正在申请“云计算”商标，此举旨在加强对这一未来可能重塑技术架构的术语的控制权。2010年3月5日，Novell与云安全联盟(CSA)共同宣布一项供应商中立计划，名为“可信任云计算计划(Trusted Cloud Initiative)”。2010年7月，美国国家航空航天局和包括Rackspace、AMD、英特尔、戴尔等支持厂商共同宣布“OpenStack”开放源代码计划，微软在2010年10月表示支持OpenStack与Windows Server 2008 R2的集成；而Ubuntu已把OpenStack加至11.04版本中。2011年2月，思科系统正式加入OpenStack，重点研制OpenStack的网络服务。2014年3月，中国国际云计算技术和应用展览会在北京开幕，云计算综合标准化技术体系已形成草案。工信部要我国从五个方面促进云计算快速发展：一是要加强规划引导和合理布局，统筹规划全国云计算基础设施建设和云计算服务产业的发展；二是要加强关键核心技术研发，创新云计算服务模式，支持超大规模云计算操作系统，核心芯片等基础技术的研发推动产业化；三是要面向具有迫切应用需求的重点领域，以大型云计算平台建设和重要行业试点示范、应用带动产业链上下游的协调发展；四是要加强网络基础设施建设；五是要加强标准体系建设，组织开展云计算以及服务的标准制定工作，构建云计算标准体系。

3.8.3 云计算对多源信息融合技术实现的影响

云计算环境下，软件开发的环境、工作模式也将发生变化。基于互联网技术应用的多源信息融合系统，在云计算环境中也必然发生很大变化。

(1) 多源信息融合系统所开发的软件，必须与云计算相适应，能够与虚拟化为核心的云平台有机结合，适应运算能力、存储能力的动态变化。虽然，传统的软件工程理论不会

发生根本性的变革，但基于云平台的开发工具、开发环境、开发平台将为敏捷开发、项目组内协同、异地开发等带来便利。软件开发项目组内可以利用云平台，实现在线开发，并通过云实现知识积累、软件复用。软件测试的环境也可移植到云平台上，通过云构建测试环境；软件测试也应该可以通过云实现协同、知识共享和测试复用。

(2) 新的软件要能够满足大量用户的使用，包括数据存储结构、处理能力。在云平台上，软件可以是一种服务，如 SaaS，也可以就是一个 Web Services，如苹果的在线商店中的应用软件等。

(3) 要互联网化，基于互联网提供软件的应用。

(4) 安全性要求更高，可以抗攻击，并能保护私有信息。

(5) 可工作于移动终端、手机、网络计算机等各种环境。

3.9 小结

本章主要给出了与多源信息融合相关的智能计算与识别理论基础，不包含已经熟知的神经元网络，以及模式识别教程中的一般内容。本章讨论的大都是二值问题，且与分类问题密切相关。首先给出了有关模式识别、智能学习与统计模式识别的一般概念，然后讲述了信息系统与粗糙集理论的基础知识。然后重点讲述了在信息融合中有特殊意义的证据理论及 D-S 合成公式，以及证据推理的一般方法。随机集理论是近年来颇受学术界关注的信息融合新工具，但其理论基础和方法还有待进一步研究。支持向量机是当前统计学习研究的热点，其优越的性能使得这一方法得到普遍重视。Bayes 网络也是一个非常有希望的工具，但实用方法仍需要进一步开发。本章内容相对分散，一般给出比较完整的表述，但难免有支离破碎的感觉，希望读者在此基础上参阅有关文献。

本章最后讲述了大数据时代的云计算及其对信息融合技术实现的影响，这是当前和未来发展的一个重要技术领域。

参考文献

[1] 韩崇昭. 应用泛函分析——自动控制的数学基础[M]. 北京：清华大学出版社，2008.

[2] Marek W，Pawlak Z. Rough sets and information systems[J]. Fundamenta Informaticae，1984，17：105-115.

[3] Pawlak Z. Rough Sets：Theoretical Aspects of Reasoning about Data[M]. Dordrecht：Kluwer Academic Publishers，1991.

[4] Pawlak Z. Rough Sets[J]. International Journal of Computer and Information Sciences. 1982，11：341-356.

[5] 张文修，梁怡，吴伟志. 信息系统与知识发现[M]. 北京：科学出版社，2003.

[6] 张文修，米据生，吴伟志. 不协调目标信息系统的知识约简[J]. 计算机学报，2003，26(1)：12-18.

[7] Yao Y Y. Generalized Rough Set Models[C]//Polkowski L，Skowron A，eds. Rough sets in Knowledge Discovery. Heidelberg：Physica-Verlag，1998. 286-318.

[8] 韩德强. 基于证据推理的多源分类融合理论与方法研究[D]. 西安：西安交通大学，2008.

[9] Klir G J，Yuan B. Fuzzy Sets and Fuzzy Logic：Theory and Applications[M]. Upper Saddle River，NJ：Prentice-Hall，1995.

[10] Hartley R V L. Transmission of information[J]. Bell System Technical Journal, 1928, 7: 535-563.

[11] Shannon C E. A mathematical theory of communication[J]. Bell System Technical Journal, 1948, 27: 379-423, 623-656.

[12] Dubois D, Prade H. A note on measures of specificity for fuzzy sets[J]. International Journal of General Systems, 1985, 10 (4): 279-283.

[13] Harmanec D, Klir G J. Measuring total uncertainty in Dempster-Shafer theory[J]. International Journal of General Systems, 1994, 22 (4): 405-419.

[14] Klir G J, Wierman M J. Uncertainy-Based Information[M]. 2nd ed. Heidelberg, Germany: Physica-Verlag, 1999.

[15] Klir G J, Smith R M. Recent developments in generalized information theory[J]. International Journal of Fuzzy Systems, 1999, 1 (1): 1-13.

[16] 戴冠中，潘泉，张山鹰，等. 证据推理的进展及存在问题[J]. 控制理论与应用，1999，(04)：465-469.

[17] 杨阳. 证据推理组合方法的分类、评价准则及应用研究[D]. 西安：西北工业大学，2006.

[18] Denoeux T. An evidence-theoretic neural network classifier [C]//Proceedings of IEEE International Conference on Systems, Man and Cybernetics, Vancouver, 1995: 712-714.

[19] Barnett J A. Computational methods for a mathematical theory of evidence[C]//Proceedings of the 7th International Joint Conference on Artificial Intelligence, Vancouver, Canada, 1981: 868-875.

[20] Schubert J. On nonspecific evidence[J]. International Journal of Intelligent Systems, 1993, 8: 711-725.

[21] 孙怀江，胡钟山，杨静宇. 学习相关源证据[J]. 南京大学学报(自然科学版)，2001，(02)：154-158.

[22] 孙怀江，杨静宇. 一种相关证据合成方法[J]. 计算机学报，1999，(09)：1004-1007.

[23] 孙怀江，胡钟山，杨静宇. 基于证据理论的多分类器融合方法研究[J]. 计算机学报，2001，(03)：231-235.

[24] Wu Y G, Yang J Y, Ke L, et al. On the evidence inference theory[J]. Information Sciences, 1996, 89 (3-4): 245-260.

[25] Smets P. The transferable belief model [J]. Artificial Intelligence, 1994, 66 (2): 191-234.

[26] Smarandache F, Dezert J. Advances and Applications of DSmT for Information Fusion Vol II [M]. Rehoboth: American Research Press, 2006.

[27] Dezert J, Smarandache F. On the generation of hyper-powersets for the DSmT[C]//Proceedings of the 6th International Conference on Information Fusion, 2003: 1118-1125.

[28] Dezert J. Foundations for a new theory of plausible and paradoxical reasoning[J]. Information and Security Journal, 2002, 9: 13-57.

[29] Matheron G. Random Sets and Integral Geometry[M]. New York: John Wiley, 1975.

[30] Molchanov I S. Limit Theorems for Union of Random Closed Sets. Lecture Notes in Mathematics[M]. Springer-Verlag, 1993.

[31] Peng T, Wang P, Kandel A. Knowledge acquisition by random sets[J]. International Journal of Intelligent Systems, 1997, 11: 113-147.

[32] Sanchez L. A random sets-based method for identifying fuzzy models[J]. Fuzzy Sets and Systems, 1998, 343-454.

[33] Nunez-Garcia J, Wolkenhauer O. Random Sets and Histograms[EB/OL]. [2010-07-20]. http://www.umist.ac.uk/csc/.

[34] Miranda E, Couso I, Gil P. Extreme points of credal sets generated by 2-alternating capacities[J]. International Journal of Approximate Reasoning, 2003, 33: 95-115.
[35] 孙怀江,杨静宇. 一种相关证据合成方法[J]. 计算机学报,1999,22(9): 272-290.
[36] 何友,王国宏,路大金,等. 多传感器信息融合及应用[M]. 北京: 电子工业出版社,2000.
[37] Selzer F, Gutfinger D. LADAR and FLIR based sensor fusion for automatic target classification [C]//Proceedings of SPIE, Montreal, Canada, 1988: 236-246.
[38] Shafer G. A Mathematical Theory of Evidence[M]. Princeton: Princeton University Press, 1967.
[39] Zadeh L A. Fuzzy sets[J]. Information Control, 1965, 8: 338-353.
[40] 谢季坚,刘承平. 模糊数学方法及其应用[M]. 2 版. 武汉: 华中科技大学出版社,2000.
[41] Bi Y X, Bell D, Wang H, et al. Combining multiple classifiers using Dempster's rule of combination for text categorization [M]//. Lecture Notes in Artificial Intelligence-Modeling Decisions for Artificial Intelligence. Berlin: Springer Verlag, 2004: 127-138.
[42] Valente F, Hermansky H. Combination of acoustic classifiers based on Dempster-Shafer theory of evidence[C]//Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing, 2007: IV-1129-IV-1132.
[43] Bi Y X, Bell D, Guan J W. Combining evidence from classifiers in text categorization [C]// Proceedings of the 8th International Conference on KES, Wellington, New Zealand 2004: 521-528.
[44] 段新生. 证据理论与决策、人工智能[M]. 北京: 中国人民大学出版社,1990.
[45] 张文修,梁怡. 不确定性推理原理[M]. 西安: 西安交通大学出版社,1996.
[46] Begler P L. Shafer-Dempster reasoning with applications to multisensor target identification system[J]. IEEE Transactions on Systems, Man and Cybernetics, 1987, 17: 968-977.
[47] Zhu Y M, Dupuis O, Kaftandjian V, et al. Automatic determination of mass functions in Dempster-Shafer theory using fuzzy c-means and spatial neighborhood information for image segmentation[J]. Optical Engineering, 2002, 41(4): 760-770.
[48] Salzenstein F, Boudraa A O. Iterative estimation of Dempster-Shafer's basic probability assignment: application to multisensor image segmentation[J]. Optical Engineering, 2004, 43 (6): 1293-1299.
[49] 孙亮. 高维遥感数据融合与分类的知识发现方法研究[D]. 西安: 西安交通大学,2006.
[50] Moyal J E. The general theory of stochastic population processes[J]. Acta Mathematica, 1962, 108: 1-31.
[51] Mathéron G. Random Sets and Integral Geometry[M]. New York: Wiley, 1975.
[52] Ripley B. Locally finite random sets: foundations for point process theory[J]. Annals of Prob, 1976, 6(4): 983-994.
[53] Baudin M. Multidimensional Point Processes and Random Closed Sets[J]. Journal of Applied Prob, 1984, 21: 173-178.
[54] Mahler R. Nonadditive Probability, Finite-set Statistics and, and Information Fusion[C]//New Orleans, LA 1995: 1947-1952.
[55] El-Fallah A, Mahler R, Ravichandrana B, et al. Adaptive Data Fusion Using Finite-Set Statistics [C]//Orlando, Florida 1999: 80-91.
[56] Mahler R. Random Sets: Unification and Computation for Information Fusion-A Retrospective Assessment[C]//Stockholm, Sweden, 2004.
[57] Mahler R P S. "Statistics 101" for Multisensor, Multitarget Data Fusion[J]. IEEE A&E SYSTEMS MAGAZINE, 2004, 19 (1): 53-64.
[58] Mahler R P S. The Random-Set Approach to Data Fusion[J]. SPIE, 1994, 2237: 287-295.
[59] Mahler R. Multisensor-multitarget statistics. Data Fusion Handbook[M]. Boca Raton: CRC

Press，2001.
[60] Mahler R. The search for tractable Bayesian multitarget filters[C]. SPIE，2000：310-320.
[61] Nguyen H T. An Introduction to Random Sets[M]. Taylor & Francis Group，LLC，2006.
[62] Goodman I R，Mahler R P S，Nguyen H T. Mathematics of data fusion[M]. Kluwer academic publishers，1997.
[63] Vo B，Singh S，Doucet A. Sequential Monte Carlo Methods for Multi-Target Filtering with Random Finite Sets[J]. IEEE Transactions on Aerospace and Electronic systems，2005，41(4)：1224-1245.
[64] Fletcher R. Practical Methods of Optimization[M]. 2nd ed. Hoboken：John Wiley and Sons，Inc.，1987.
[65] Christopher J C Burges. A Tutorial on Support Vector Machines for Pattern Recognition[J]. Data Mining and Knowledge Discovery，1998，2：121-167.
[66] Vapnik V. Estimation of Dependences Based on Empirical Data [M]. Moscow：Nauka，1979 (English translation：New York：Springer Verlag，1982).
[67] Vapnik V. The Nature of Statistical Learning Theory[M]. New York：Springer Verlag，1995.
[68] Vapnik V，Golowich S，Smola A. Support vector method for function approximation，regression estimation and signal processing[J]. Advances in Neural Information Processing Systems，1996，9：281-287.
[69] Strang G T. Introduction to Applied Mathematics [M]. Wellesley：Wellesley-Cambridge Press，1986.
[70] Scholkopf B，Sung K，Burges C，et al. A comparing support vector machines with Gaussian kernels to radial basis function classifiers[J]. IEEE Trans. Sign. Processing，1997，45：2758-2765.
[71] Suykens J A K，Vandewalle J，De Moor B. Optimal Control by Least Squares Support Vector Machines[J]. Neural Networks，2001，14(1)：23-35.
[72] Zhu J Y，Ren B，Zhang H X，et，al. Time Series Prediction via New Support Vector Machines[J]. IEEE，In Proceeding of 2002 ICMLC，China Beijing，2002.
[73] Lin C F，Wang S D. Fuzzy Support Vector Machine[J]. IEEE Trans. On Neural Networks，2002，13(2)：28-29.
[74] Zhang L，Zhou W D，Jiao L C. wavelet support vector machine[J]. IEEE Trans. On SMC-Part B：CYBERNETICS，2004，34(1)：120-125.
[75] Zhang Q H，Benveniste A. Wavelet networks[J]. IEEE Trans. Neural Networks，1992，3：889-898.
[76] Szu H H，Telfer B，Kadambe S. Neural network adaptive wavelets for signal representational and classification[J]. Opt. Eng.，1992，31：1907-1916.
[77] 张恒喜，郭基联，朱家元，等. 小样本多元数据分析方法及应用[M]. 西安：西北工业大学出版社，2002.
[78] 肖健华，吴今培. 基于核的特征提取技术及应用研究[J]. 计算机工程，2002，28(10)：36-38.
[79] Lima A，Zen H，Nankaku Y，et al. On the Use of Kernel PCA for Feature Extraction in Speech Recognition[J]. IEICE Transactions on information and systems，2004，E87-D(12)：2802-2811.
[80] Bernhard S，John C P. Estimating the Support of a High-Dimensional Distribution[R]. Microsoft Research Technical Report MSR-TR-99-87，1999，1-28.
[81] Scholkopf B. Statistical Learning and Kernel Methods[R]. Microsoft Research Technical Report MSR-TR-2000-23，2000，1-29.
[82] Suykens J A K，Vandewalle J，De Moor B. Optimal Control by Least Squares Support Vector Machines[J]. Neural Networks，2001，14(1)：23-35.
[83] Charniak E. Bayesian Networks without Tears[EB/OL]. [2010-07-21]. http://www.aaai.org.
[84] Bottcher S G，Dethlefsen C. Learning Bayesian Networks with R[C]//Proceedings of the 3rd

International Workshop on Distributed Statistical Computing (DSC) March 20-22, Vienna, Austria,2003.

[85] Koivisto M,Sood K. Exact Bayesian Structure Discovery in Bayesian Networks[J]. Journal of Machine Learning Research 5,2004,549-573.

[86] Tessem B. Approximations for efficient computation in the theory of evidence[J]. Artificial Intelligence, 1993, 61(2): 315-329.

[87] Smets P. Decision making in the TBM: The necessity of the pignistic transformation[J]. International Journal of Approximate Reasoning, 2005, 38(2): 133-147.

[88] Han D Q, Dezert J, Han C Z, et al. New dissimilarity measures in evidence theory[C]// Proceedings of the 14th International Conference on Information Fusion, Chicago, IL, USA, 2011: 1-7.

[89] Jousselme A L, Grenier D, Bossé É. A new distance between two bodies of evidence[J]. Information Fusion, 2001, 2(2): 91-101.

[90] Han DQ, Dezert J, Yang Y. Belief Interval-Based Distance Measures in the Theory of Belief Functions[J]. IEEE Transactions on Systems, Man, and Cybernetics: Systems, 2018, 48(6): 833-850.

[91] Irpino A, Verde R. Dynamic clustering of interval data using a Wasserstein-based distance[J]. Pattern Recognition Letters, 2008, 29(11): 1648-1658.

[92] Vo B T, Vo B N. A random finite set conjugate prior and application to multi-target tracking [C]//2011 Seventh International Conference on Intelligent Sensors, Sensor Networks and Information Processing. IEEE, 2011: 431-436.

[93] Vo B T, Vo B N. Labeled random finite sets and multi-object conjugate priors[J]. IEEE Transactions on Signal Processing, 2013, 61(13): 3460-347.

[94] Höhle U. Entropy with respect to plausibility measures[C]//Proceedings of the 12th IEEE International Symposium on Multiple-Valued Logic, 1982: 167-169.

[95] Yager RR. Entropy and specificity in a mathematical theory of evidence[J]. International Journal of General Systems, 1983, 9(4): 249-260.

[96] Klir G J, Ramer A. Uncertainty in the Dempster-Shafer theory: a critical reexamination[J]. Int. J. Gen. Syst., 1990, 18(2): 155-166.

[97] Klir GJ, Parviz B. A note on the measure of discord[C]//Proceedings of the Eighth International Conference on Uncertainty in Artificial Intelligence, Morgan Kaufmann Publishers Inc., 1992: 138-141.

[98] Dubois D, Prade H. A note on measures of specificity for fuzzy sets[J]. International Journal of General Systems, 1985, 10(4): 279-283.

[99] Korner R, Nather W. On the specificity of evidences[J]. Fuzzy Sets and Systems, 1995, 71(2): 183-196.

[100] Harmanec D, Klir GJ. Measuring total uncertainty in Dempster-Shafer theory: a novel approach [J]. International Journal of General Systems, 1994, 22(4): 405-419.

[101] Jousselme A L, Liu C S, Grenier D, et al. Measuring ambiguity in the evidence theory[J]. IEEE Transactions on Systems Man and Cybernetics Part A: Systems and Humans, 2006, 36 (5): 890-903.

[102] Smets P. Decision making in the TBM: The necessity of the pignistic transformation[J]. International Journal of Approximate Reasoning, 2005, 38(2): 133-147.

[103] Yang Y, Han D Q. A new distance-based total uncertainty measure in the theory of belief functions[J]. Knowledge-Based Systems, 2016, 94: 114-123.

[104] Irpino A, Verde R. Dynamic clustering of interval data using a Wasserstein-based distance[J]. Pattern Recognition Letters, 2008, 29(11): 1648-1658.