

论正态云模型的普适性

李德毅¹, 刘常昱^{2,3}

(1. 北京电子系统工程研究所, 北京 100840; 2. 解放军理工大学, 南京 210007;
3. 武汉大学软件工程国家重点实验室, 武汉 430072)

[摘要] 分布函数是分析随机现象的重要工具, 正态分布是最重要的概率分布, 在自然科学和社会科学中应用广泛; 隶属函数是模糊集合的基石, 正态隶属函数也有广泛的应用。但是, 精确确定一个模糊概念的隶属函数已经成为模糊学应用的瓶颈。云模型把随机性和模糊性结合起来, 用数字特征熵, 揭示随机性与模糊性的关联性, 并用来表示一个定性概念的粒度。正态云模型通过期望、熵和超熵构成的特定结构发生器, 生成定性概念的定量转换值, 体现概念的不确定性。这种特定结构不但放宽了形成正态分布的前提条件, 而且把精确确定隶属函数放宽到构造正态隶属度分布的期望函数, 因而更具有普遍适用性, 更简单、直接地完成了定性定量之间的相互转换过程。

[关键词] 正态分布函数; 隶属度; 正态云模型; 概念粒度; 普适性

[中图分类号] TP182 **[文献标识码]** A **[文章编号]** 1009-1742(2004)08-0028-07

1 引言

具有不确定性的现象与事物, 普遍存在于自然界和人类社会。如何表示和处理不确定性现象与事物, 一直是自然科学研究的热点和重点, 也是一个瓶颈性问题^[1]。在各类不确定性中, 随机性、模糊性最为关键, 受到了人们更多的关注。人类的认知过程, 是通过语言和思维进行的, 可以认为是对客观世界的一种镜像。自然语言是人类的思维基础, 是人类智能的体现, 概念可以认为是自然语言和思维的基本单位, 自然语言的最小单元是语言值, 对应的概念可以认为是人类思维的基本细胞^[2]。

人类依托语言进行思维。自然语言中的许多概念, 也是不精确的、模糊的、随机的, 有时甚至是不完全的。这些不确定概念, 难以给出很精确的定义。因此, 如何表示和处理不确定性知识, 一直是人工智能研究的热点和重点, 要使计算机具有类似于人类的理解、判断能力, 就无法回避对自然语言

中概念的随机性和模糊性的研究, 尤其是这种不确定性知识的形式化表示问题。研究随机现象的主要工具是概率理论, 概率论有着坚实的理论基础, 能够很好表示和处理随机不确定^[3]。而自从 Zadeh 博士 1965 年发表了模糊集合的开创性论文后^[4], 模糊集合理论成为处理模糊不确定的主要工具, 在理论和应用中已取得有目共睹的诸多成就^[5]。云模型是在这 2 种理论进行交叉渗透的基础上, 通过特定构造的算法, 形成定性概念与其定量表示之间的转换模型, 并揭示随机性和模糊性的内在关联性。正态云模型是云模型中的一种, 具有普遍适应性。

2 正态分布的普适性与产生条件

研究随机现象的最基本工具是随机变量, 分布函数是随机变量最重要的概率特征, 它可以完全描绘随机变量的统计规律。在概率论与随机过程的理论研究和实际应用中, 正态分布起着特别重要的作用, 在各种概率分布中居于首要的地位。正态分布函数形式为

[收稿日期] 2004-04-22; 修回日期 2004-05-31

[基金项目] 国家自然科学基金资助项目 (60375016); 国家自然科学基金重大项目 (60496323)

[作者简介] 李德毅 (1944-), 男, 江苏泰县人, 中国工程院院士, 北京电子系统工程研究所研究员, 博士生导师

$$F(x, \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma}} \int_{-\infty}^x \exp\left(-\frac{(u-\mu)^2}{2\sigma^2}\right) du.$$

概率密度函数为

$$f(x, \mu, \sigma^2) = (2\pi)^{-1/2} \sigma^{-1} \exp[-(x-\mu)^2/2\sigma^2],$$

其中, μ 和 σ^2 分别是正态分布的期望和方差, 分别表征随机变量的最可能取值以及一切可能取值的分散程度。

正态分布广泛存在于自然现象、社会现象、科学技术以及生产活动中, 在实际中遇到的许多随机现象都服从或者近似服从正态分布, 例如, 正常生产条件下的产品质量指标, 随机测量误差, 同一生物群体的某种特征, 某地的年平均气温等等。中心极限定理从理论上阐述了产生正态分布的条件^[3]。对中心极限定理的简单直观说明, 如果决定某一随机变量结果的是大量微小的、独立的随机因素之和, 并且每一因素的单独作用相对均匀的小, 没有一种因素可起到压倒一切的主导作用, 那么这个随机变量一般近似于正态分布。例如, 某种成批生产的产品, 如果工艺、设备、技术、操作、原料等生产条件正常且稳定, 那么产品的质量指标应该近似服从正态分布; 否则说明生产条件不稳定或发生变化, 影响了产品质量。在实际应用中, 人们多是根据上述考虑来判断随机现象是否服从正态分布的。

正态分布是许多重要概率分布的极限分布, 许多非正态的随机变量是正态随机变量的函数, 正态分布的密度函数和分布函数有各种很好的性质和比较简单的数学形式, 这些都使得正态分布在理论和实际中应用非常广泛。

在强调正态分布的地位同时, 必须指出许多随机现象不能用正态分布来描绘。如果决定随机现象的因素单独作用不是均匀的小, 相互之间并不独立, 有一定程度的相互依赖, 就不能够符合正态分布的产生条件, 不构成正态分布, 或者只能用正态分布来近似处理。概率论用联合分布来处理这类情况, 但是通常联合概率分布的确定非常复杂, 难以实际应用。笔者提出用云模型来描述这类随机性, 将正态分布扩展为泛正态, 用一个新的独立参数——超熵, 来衡量偏离正态分布的程度, 这种处理方法比单纯用正态条件分布更为宽松, 同时比联合分布简单, 易于表示和操作。

3 模糊集和正态隶属函数

隶属函数是模糊集合理论的基石, 它是度量模

糊程度的函数。通过隶属函数, 模糊理论将模糊现象转变成精确数学加以研究。从模糊集合的开创者 Zadeh 开始, 近 40 年来, 人们普遍接受并使用的模糊集合及隶属函数定义, 可表述如下^[6]:

定义 设 X 是一个普通集合 $X = \{x\}$, 称为论域。在论域 X 上的一个模糊集合 \tilde{A} 定义为一个隶属函数 $\mu_{\tilde{A}}(x): X \rightarrow [0, 1]$, $x \in X$, 它把论域 X 中的元素映射到 $[0, 1]$ 中的实数。其中 $\mu_{\tilde{A}}(x)$ 称为元素 x 隶属于模糊集合 \tilde{A} 的程度, 简称为 x 对 \tilde{A} 的隶属度。隶属函数 $\mu_{\tilde{A}}(x)$ 是论域中所有元素属于模糊集合 \tilde{A} 的隶属度分布。模糊集合就是一个有着不同隶属度的元素集合。

自然和社会科学中的大量模糊概念的隶属函数, 并没有严格的确定方法, 通常靠经验确定, 归纳起来大致有 6 种形态, 为简便起见, 这里仅取它们的简化解析形式:

- 1) 线性隶属函数 $\mu_{\tilde{A}}(x) = 1 - kx$;
- 2) Γ 隶属函数 $\mu_{\tilde{A}}(x) = e^{-kx}$;
- 3) 凹 (凸) 形隶属函数 $\mu_{\tilde{A}}(x) = 1 - ax^k$;
- 4) 柯西隶属函数 $\mu_{\tilde{A}}(x) = 1/(1 + kx^2)$;
- 5) 岭形隶属函数 $\mu_{\tilde{A}}(x) = 1/2 - (1/2)\sin[(\pi/(b-a))(x - (b-a)/2)]$;
- 6) 正态隶属函数 $\mu_{\tilde{A}}(x) = \exp[-(x-a)^2/2b^2]$ ^[6]。

在描述模糊概念的亦此亦彼性的时候, 前 3 种隶属函数形态在亦此亦彼性的刻画上, 虽然连续, 但出现突变点, 即函数曲线的一阶导数不连续, 这种突变不符合中介过渡性质的渐变特征。如果认为模糊性在宏观和微观上都存在, 概念在不同尺度上都连续, 则高阶导数也应该连续。另外, 隶属函数在靠近“亦此”或“亦彼”区间的变化率反而较大, 也不符合一般的认识规律。因此这 3 种隶属函数仅用在一些简单的场合。

对柯西、岭形、正态 3 种函数形态的比较分析:

中国学者张南伦等人^[6]曾经用统计方法对模糊概念“青年人”求得的隶属度为例进行仔细分析, 这是一个典型的求隶属度的大型试验。张在武汉建材学院选取了 129 位合适人选, 让他们独立给出各自认为的“青年人”的最适宜年龄段, 然后分组计算相对频度, 每组以中值为代表计算隶属频率, 即为各点的隶属度, 其结果如表 1 所示。

根据表 1 数据拟合出“青年”隶属函数的多项

表1 “青年”的隶属度

Table 1 The membership degree of the notion “young”

年龄 X_i	24	25	26	27	28	29	30
隶属度 Y_i	1	0.9922	0.7981	0.7829	0.7674	0.6202	0.5969
年龄 X_i	31	32	33	34	35	36	37
隶属度 Y_i	0.2093	0.2093	0.2016	0.2016	0.2016	0.0078	0

式曲线为

$$\mu_0(x) = 1.01302 - 0.00535(x - 24) - 0.00872(x - 24)^2 + 0.0005698(x - 24)^3,$$

分别用柯西型、岭形、正态隶属函数拟合 $\mu_0(x)$, 使得目标函数

$$\left(\int_a^b [\mu(x) - \mu_0(x)]^2 dx / (b - a) \right)^{1/2}$$

最小, 得到的柯西型隶属函数为

$$\mu_1(x) = 1/[1 + (x - 24)^2/30],$$

岭形隶属函数为

$$\mu_2(x) = 1/2 -$$

$$(1/2)\sin[(\pi/(37 - 24))(x - (37 - 24)/2)],$$

正态隶属函数为

$$\mu_3(x) = \exp[-9(x - 24)^2/338].$$

通过数值积分计算这3个隶属函数和 $\mu_0(x)$ 的均方差, 共采样1400个点, 计算结果见表2, 即3种隶属函数解析式和拟合曲线 $\mu_0(x)$ 比较, 均方差最小的是正态隶属函数。

表2 不同形式隶属函数与 $\mu_0(x)$ 的均方差

Table 2 Mean variances of different membership functions and $\mu_0(x)$

隶属函数形式	与 $\mu_0(x)$ 的均方差
柯西形	0.042 181 118 428 255 4
岭形	0.060 183 795 103 931
正态形	0.030 915 588 518 457

若用这3种隶属函数和测试值 (X_i, Y_i) 进行差值比较, 计算均方差的结果见表3, 即这3种隶属函数和测试值比较, 均方差最小的仍然是正态隶属函数。

这一结果表明, 对大量模糊概念, 用正态隶属函数刻画最适合, 最接近人类思维。实际上, 人们对搜集到的各类隶属函数进行分析发现, 许多领域的隶属函数都和正态隶属函数有相当的一致性, 并

且大多数都是正态隶属函数泰勒展开式的若干低次项之和, 是正态隶属函数的一种近似。因此有理由相信, 相对于其他类型函数, 正态隶属函数在众多领域有着最广泛的应用。

表3 不同形式隶属函数与 (X_i, Y_i) 的均方差

Table 3 Mean variances of different membership functions and (X_i, Y_i)

隶属函数形式	均方差
多项式拟合曲线	0.083 742 138 813 298 3
柯西形	0.101 253 101 730 926
岭形	0.095 237 099 589 215 1
正态形	0.080 769 137 104 025

在精确的隶属函数确定之后, 就可以计算论域中的值属于这个模糊概念的隶属度, 这是一个唯一的、精确的数值。这种用精确的隶属函数来严格表示模糊概念, 得到完全确定的、清晰的关系, 本身违背了模糊学的基本精神, 已经完全被纳入精确数学范畴而没有丝毫的模糊性可言。因此, 隶属度的分布特性和分布函数受到越来越多地研究, 允许隶属度在一个中心值附近做微小摆动, 即将精确隶属度用一个有稳定倾向的随机数来代替, 将精确隶属函数用一个有稳定倾向的期望隶属曲线表示, 成为研究不确定性知识表示的重要方法。

尽管随机性和模糊性是从2种不同出发点研究引出的不确定性, 表示可能性的隶属度和表示随机性的概率分别出自不同的学科分支, 但是采用统计的方法确定隶属函数已经成为事实上的重要方法。于是, 研究人类认知过程中模糊性和随机性之间存在的关联显得十分必要。假设取能够代表一个模糊概念的样本点集, 那么具有较大隶属度的样本显然应该有较大的获取概率, 传统的隶属函数无法反映出它与随机性的这种关联。

4 定性定量不确定转换模型——正态云模型

正态云模型是用语言值表示的某个定性概念与其定量表示之间的不确定性转换模型, 它主要反映客观世界中事物或人类知识中概念的2种不确定性: 模糊性(边界的亦此亦彼性)和随机性(发生的概率), 并把二者完全集成在一起, 构成定性和定量相互间的映射, 研究自然语言中的最基本语言值(又称语言原子)所蕴含的不确定性的普遍规

律,使得有可能从语言值表达的定性信息中获得定量数据的范围和分布规律,也有可能把精确数值有效转换为恰当的定性语言值。

正态云模型用相互独立的一组参数共同表达一个定性概念的数字特征,反映概念的不确定性。在正态分布函数与正态隶属函数基础上,这组参数用期望 E_x , 熵 E_n , 超熵 H_e 这3个数字特征来表征:

期望 E_x 在论域空间中最能够代表这个定性概念的点,是这个概念量化的最典型样本点。

熵 E_n 代表一个定性概念的可度量粒度,通常熵越大概念越宏观。熵还反映了定性概念的不确定性,表示在论域空间可以被定性概念接受的取值范围大小,即模糊度,是定性概念亦此亦彼性的度量。

超熵 H_e 熵的不确定性的度量,它反映代表性概念值的样本出现的随机性,揭示了模糊性和随机性的关联。

当概念对应的论域为一维时,定性定量的不确定性转换通过正态云发生器算法实现。具体算法如下^[7]:

输入 表示定性概念 \tilde{A} 的3个数字特征值 E_x, E_n, H_e , 云滴数 N 。

输出 N 个云滴的定量值,以及每个云滴代表概念 \tilde{A} 的确定度。

算法:

Step 1 生成以 E_n 为期望值, H_e 为标准差的一个正态随机数 E'_n ;

Step 2 生成以 E_x 为期望值, $\text{abs}(E'_n)$ 为标准差的正态随机数 x ;

Step 3 令 x 为定性概念 \tilde{A} 的一次具体量化值,称为云滴;

Step 4 计算 $y = \exp[-(x - E_x)^2 / 2(E'_n)^2]$;

Step 5 令 y 为 x 属于定性概念 \tilde{A} 的确定度;

Step 6 $\{x, y\}$ 完整地反映了这一次定性定量转换的全部内容;

Step 7 重复 Step 1 至 Step 6,直到产生 N 个云滴。

从这个特定的生成结构序可以看出,正态云模型充分运用正态分布和正态隶属函数的普适性,进一步放宽了约束条件,简化了参数,并将随机性与模糊性通过熵统一表示,构成不确定性知识表示模型。

对于一个严格意义正态分布的随机变量,如果期望已知,正态方差的共轭先验分布为倒伽玛分布,这种先验分布显示了期望与方差的相关性,方差的变化会对期望产生影响^[8]。在云发生器构成过程中,把熵 E_n 看成是与期望 E_x 独立的另一个数字特征,同时考虑到正态分布的普适性,又把熵 E_n 作为超熵 H_e 的期望,用正态分布产生随机熵 E'_n 。再用随机熵 E'_n 作为期望 E_x 的熵,生成不确定性概念的一次正态分布的随机精确量值。这里不是在讨论单一随机变量的概率分布,而是设计一个特定结构算法,利用方差的随机性创造一种崭新的随机变量形态,利用这一随机变量来表示不确定性概念。下一节将证明云发生器算法中的 X 给出的是泛正态分布。

在数域空间,正态云模型既不是一个确定的概率密度函数,也不是一条明晰的隶属函数曲线,而是由2次串接的正态发生器生成的许多云滴组成的、一对多的泛正态数学映射图像,是一朵可以伸缩、无确定边沿、有弹性的云图,完成定性和定量之间的相互映射。

通常用期望曲线方法研究数据集在空间随机分布的统计规律性,云模型无确定边沿但是有整体形状,笔者用云 (X, Y) 的期望曲线来刻画云模型的轮廓特征。定义函数 $y = \exp[-(x - E_x)^2 / 2(E_n)^2]$ 为正态云 (X, Y) 的期望曲线。期望曲线光滑的穿过云滴“中间”,它勾画出云的整体“轮廓”,是云滴集合的“骨架”,所有的云滴都在期望曲线附近做随机的波动,而波动程度的大小是由 H_e 来控制的。这里的“中间”并不是几何意义的中间,而是概率意义下的“中间”——期望。

云模型不再强调精确的函数表示,而是利用3个数字特征表示概念的不确定性,通过特定的计算机算法来实现定性概念和定量表示的不确定转换,同时揭示了模糊性和随机性之间的关联。云模型勾勒出不确定概念的轮廓,做出近似的、灵活性的结论。众所周知,适当的模糊反而“精确”,过分的追求精确反而“模糊”。这正像人们平常的谈话和思维,常常使用不精确的,含糊的术语,但是并不妨碍人们正确理解其所表达的内容,也不妨碍人们得出正确结论。云模型仅仅用3个数字特征和这个特定算法,将语言值量化了。

在人的意识活动中,概念是可以在不同尺度空间转移和提升的,不同尺度空间的概念具有不同的

粒度。概念的粒度可以用云模型中的熵 E_n 度量。由于云模型是泛正态的, 约束条件比正态分布还宽松, 因此云模型可以表示大量宏观、中观、微观不同尺度上的概念, 能够反映概念对应的客观事物的粒度。 E_n 越大, 对应概念粒度越大, 概念越宏观^[9]。

云模型中的3个数字特征可以是多维的, 表示一个概念的多维属性。其算法只要把2次串接的正态发生器扩展到串接的多维正态发生器即可。

云模型中的期望还可以扩展为一个图形、一幅图像、一个签名、一段声音等等, 扩展云模型可以给出这个图形、这幅图像、这个签名、或这段声音任意数目的扩展图像、扩展签名、扩展声音、扩展数码, 这种扩展带有不确定性。通过扩展云可以表示更广泛, 更复杂的不确定知识。

5 正态云模型的普适性证明

5.1 云 X 的概率密度

由正态云发生器算法可知, E'_n 服从以 E_n 为期望、 H_c^2 为方差的正态分布, 所以 E'_n 的概率密度为

$$f_{E'_n}(x) = (2\pi)^{-1/2} H_c^{-1} \exp[-(x - E_n)^2 / 2H_c^2];$$

而 X 服从以 E_x 为期望、 E_n^2 为方差的正态分布, 此时 X 的概率密度为

$$f_X(x) = (2\pi)^{-1/2} |E'_n|^{-1} \exp[-(x - E_x)^2 / 2E_n^2];$$

根据条件密度函数公式, X 的概率密度函数为

$$f_X(x) = \int_{-\infty}^{+\infty} \frac{1}{2\pi H_c |y|} \cdot \exp\left[-\frac{(x - E_x)^2}{2y^2} - \frac{(y - E_n)^2}{2H_c^2}\right] dy.$$

这是一个没有解析形式的密度函数, 其分布呈现“中间多, 两头少”的特点, 当 $H_c = 0$ 时, 该函数是正态分布 $N(E_x, E_n^2)$ 的密度函数。

由正态云的算法可知: X 的期望为

$$E(X) = E_x,$$

方差为

$$D(X) = \int_{-\infty}^{+\infty} (x - E_x)^2 dx \int_{-\infty}^{+\infty} \frac{1}{2\pi H_c |y|} \cdot \exp\left[-\frac{(x - E_x)^2}{2y^2} - \frac{(y - E_n)^2}{2H_c^2}\right] dy = \frac{1}{\sqrt{2\pi} H_c} \int_{-\infty}^{+\infty} |y| \exp\left[-\frac{(y - E_n)^2}{2H_c^2}\right] dy.$$

$$\int_{-\infty}^{+\infty} \frac{1}{\sqrt{2\pi}} \frac{(x - E_x)^2}{y^2} \exp\left[-\frac{(x - E_x)^2}{2y^2}\right] dx = \frac{1}{\sqrt{2\pi} H_c} \int_{-\infty}^{+\infty} y^2 \exp\left[-\frac{(y - E_n)^2}{2H_c^2}\right] dy \cdot \int_{-\infty}^{+\infty} \frac{1}{\sqrt{2\pi}} t^2 \exp\left[-\frac{t^2}{2}\right] dt = E_n^2 + H_c^2$$

公式隐含的意义如下: 通常认为, 如果某一现象决定于若干独立的、微小的随机因素的总和, 并且各个因素的单独作用相对均匀地小, 那么这一现象近似于正态分布。但是, 在很多情况下, 导致随机现象产生的多种不确定因素未必是严格均匀的, 也未必是完全独立的, 如果用正态分布来简单近似, 就不能真实反映客观情况。例如, 枪手射击, 若每一次射击都是独立的, 不考虑外界因素对射手的影响, 靶标上靶点的分布可用正态分布来表示。但是在实际射击过程中, 已经射出的成绩会对下一次有影响, 枪手的心理素质会对射击结果有比较大的影响, 尤其是在比赛等特殊场合。对此, 云模型比正态分布更合理: 可以利用超熵 H_c 来反映心理因素。心理素质好的, H_c 就较小, 这时 $E_n^2 + H_c^2$ 和 E_n^2 差距不大, 整体的射击结果就比较稳定, 接近于其正常的射击水平; 而如果射击者的心理素质较差, H_c 就较大, $E_n^2 + H_c^2$ 和 E_n^2 的差距也大, 反映出打靶结果的不稳定性, 同射击者正常的射击水平相差较大。

上述结论证明, 云模型中的参数 H_c 可以反映影响因素中的不均匀情况或者不相互独立的情况, 是偏离正态分布程度的度量。称云 X 的这种分布为泛正态分布。泛正态分布有其产生的实际背景和条件: 如果影响某一随机变量结果的因素不完全独立, 或者某几种因素作用突出, 那么这一随机变量服从泛正态分布。泛正态分布的产生条件没有正态分布苛刻, 在实际中, 为简便, 我们将许多情况近似当成正态分布来处理, 但未必符合正态分布的产生条件。实际上泛正态更接近客观实际, 它比正态分布复杂但是形成条件远比正态分布宽松, 又远比联合分布简便、可行。云模型是可以生成泛正态分布的算法, 同时给出带有随机性的隶属度值, 不宜简单地看成是二次正态或者二次随机。

泛正态的普适意义比正态分布更广, 当 $H_c = 0$ 时云 X 退化为正态分布。

5.2 云 Y ——不确定中的规律性

根据上述算法产生的云模型的确度 y 是具

有固定概率密度的随机变量，而与云模型的 3 个参数无关，证明如下：

根据算法，将 y_i 看作是随机变量 $Y_i = \exp[-(X - E_x)^2/2(E'_{ni})^2]$ 的一个样本，其中 E'_{ni} 是 E'_n 的一次实现。先求出 Y_i 的分布函数 $F_{Y_i}(y)$ 。当 $y \in (0, 1)$ 时，

$$\begin{aligned} F_{Y_i}(y) &= P\{Y_i \leq y\} = \\ &P\{\exp[-(X - E_x)^2/2(E'_{ni})^2] \leq y\} = \\ &1 - P\{-(-2 \ln y)^{1/2} \leq \\ &(X - E_x)/E'_{ni} \leq (-2 \ln y)^{1/2}\}。 \end{aligned}$$

由于 X 服从正态分布 $N(E_x, E'_{ni}{}^2)$ ，所以 $(X - E_x)/E'_{ni}$ 服从标准正态分布 $N(0, 1)$ ，故

$$F_{Y_i}(y) = 1 - \int_{-\sqrt{-2 \ln y}}^{\sqrt{-2 \ln y}} (2\pi)^{-1/2} \exp(-t^2/2) dt。$$

此时， Y_i 的概率密度为

$$\begin{aligned} f_{Y_i}(y) &= F_{Y_i}(y)' = \\ &(2\pi)^{-1/2} e^{-(-2 \ln y)/2} [(-(-2 \ln y)^{1/2})' - \\ &((-2 \ln y)^{1/2})'] = (-\pi \ln y)^{-1/2} \quad y \in (0, 1)。 \end{aligned}$$

当 $y \geq 1$ 时， $F_{Y_i}(y) = 1$ ；

当 $y \leq 0$ 时， $F_{Y_i}(y) = 0$ ；

故 Y_i 的概率密度为

$$f(y) = \begin{cases} (-\pi \ln y)^{-1/2}, & 0 < y < 1 \\ 0, & \text{else.} \end{cases}$$

可以看出，无论 E'_{ni} 取何值，随机变量 Y_i 的密度函数都不变，即所有的 y_i 都来自密度为

$$f(y) = \begin{cases} (-\pi \ln y)^{-1/2}, & 0 < y < 1 \\ 0, & \text{else.} \end{cases}$$

的随机变量，故 $f(y)$ 就是 Y 的概率密度^[3]。

这一概率密度函数形式简单，可以很方便地分析 Y 的数学性质。例如这是一个 $(0, 1)$ 区间内单调递增的凹函数，期望 $E_Y = 2^{-1/2}$ ，方差 $D_Y = 3^{-1/2} - 1/2$ ，各阶矩为 $E_Y^n = (n+1)^{-1/2} (n \geq 1)$ 。

Y 的概率密度同云的 3 个数字特征无关，抽象出人们认知过程中的这样一个深层规律：对于特定语言值表示的任何定性概念，只要它能用正态隶属函数来近似刻画，如“青年”、“中等个”、“大概 30”等等，尽管它们各自有不同的含义，尽管不同的量化值——云滴在 X 数域的分布和物理意义会有所不同，也尽管所有云滴表现出不同的确定度，不同的人会有不同的认识，不同的时期也会有不同的认识，但是云滴确定度的统计分布，总体上

都是一个统一的形态。说明对用语言值表示的任何概念，抛弃它们的具体物理含义，它们反映在人们脑海中的认知规律是一致的，认知的不确定性中仍然有着确定的规律性，揭示了人们用不同语言值表示不同定性概念之间存在的认识上的共同性。

当 $H_c = 0$ 时，确定度 y 退化正态云模型的期望曲线——正态隶属函数。

6 结语

科学需要重复，不能重复的一次性现象，科学中一般不予研究；科学需要精确，不能量化表示的现象，很难找到数学工具去研究；提出的云模型，为克服上述困难找到一个数学方法。这个方法放松了人们在研究符合某个概率分布时的前提条件，因为这些前提条件是难以满足、甚至是难以确切叙述的；这个方法也放松了人们在研究模糊现象时人为确定隶属度的尴尬，因为找到了独立于任何具体概念的隶属度的概率密度分布规律。

正态分布的普遍性与正态隶属函数的普遍性，共同奠定了正态云模型普适性的理论基础。云 X 的泛正态分布，具有比正态分布更宽松的前提条件要求，使得正态分布成为其特例，因而具有比正态分布更广的普适性；云 Y 实现了隶属度取值的随机性，正态隶属函数成为隶属度的期望曲线，与概念特定的物理意义无关，更能反映人类思维的灵活性与规律性。距离期望 E_x 越近的云滴越集中，反映人们对概念的认知越统一；距离 E_x 越远的云滴越离散、越稀疏，反映出人们对概念的认知越不稳定、不统一。 E_n 成为概念的粒度的度量参数， H_c 揭示了随机性和模糊性的关联性。当 $H_c = 0$ 时，云 X 退化为正态分布，而云 Y 退化为正态隶属函数。

任何一种新思想都不是个别人的突发奇想，除了有其深刻的客观实践根源外，还需要先驱者们长期的积累和孕育。1923 年，著名的英国逻辑学家伯特兰·罗素首先向传统思想发难，他在“论含混性”^[10]一文中指出：“传统逻辑都习惯于假定使用的是精确的符号，它不适用于尘世生活，而仅仅适用于天国”。罗素是现代数理逻辑的完成者，他对盲目崇拜精确性的批评，意义是深远的，“认为模糊认识必定是靠不住的，这种看法是大错特错了。正好相反，模糊认识可能比精确认识更真实，因为有更多的潜在的事实能证明模糊认识。”到 1937

年, 布兰克以相同的题目著文, 进一步探讨了含混性。布兰克提出的“轮廓的一致性”(Consistency Profiles)的概念^[11], 可以看作是隶属函数的原始形态, 这是布兰克对模糊学的一个有意义的贡献。1965年系统科学家 L. A. Zadeh 提出了隶属度、隶属函数、模糊集合等基本概念, 1982年波兰数学家 Z. Pawlak 又提出粗糙集思想^[12]。100多年来统计学和40年来模糊学的蓬勃应用, 使得我们今天又在更高的层次上认识到了这种“轮廓的一致性”, 即正态云模型的普遍适用性。今天我们可以说, 认为定性认识必定是靠不住的, 这种看法是大错特错了。正好相反, 定性认识可能比定量认识更真实, 因为有更多的潜在的事实能证明定性认识。

实际上, 探讨不确定性概念和知识的表示方式, 不仅仅是人工智能领域的工作, 还需要认知科学家、计算机科学、数学、控制学、语言学、心理学甚至哲学等领域的诸多专家, 包括工程技术人员的共同努力。

参考文献

- [1] 李德毅. 知识表示中的不确定性 [J]. 中国工程科学, 2000, 2(10): 73~79
- [2] 李德毅, 淦文燕, 刘璐莹. 人工智能与认知物理学 [A]. 中国人工智能进展 2003 [C]. 北京: 北京邮电大学出版社, 2003. 6~14
- [3] 王梓坤. 概率论基础及其应用 [M]. 北京: 北京师范大学出版社, 1996
- [4] Zadeh L. A. Fuzzy sets [J]. Information and Control, 1965, (8): 338~353
- [5] Ross T. J. Fuzzy Logic with Engineering Applications [M]. Beijing: Publishing House of Electronics Industry, 2001
- [6] 李洪兴, 汪培庄. 模糊数学 [M]. 北京: 国防工业出版社, 1994
- [7] 李德毅, 孟海军, 史雪梅. 隶属云和隶属云发生器 [J]. 计算机研究和发展, 1995, 32(6): 16~21
- [8] 茆诗松主编. 统计手册 [M]. 北京: 科学出版社, 2003
- [9] 李德毅, 王 晔, 吕辉军. 知识发现机理研究 [A]. 中国人工智能进展 2001 [C]. 北京: 北京邮电大学出版社, 2001. 314~325
- [10] Russell B. Vagueness [J]. Australasian Journal of Psychology and Philosophy, 1923, (1): 84~92
- [11] Black M. Vagueness [J]. Philosophy of Science, 1937, (4): 422~431
- [12] Pawlak Z. Rough sets [J]. International Journal of Computer and Information Science, 1982, (11): 341~356

Study on the Universality of the Normal Cloud Model

Li Deyi¹, Liu Changyu^{2,3}

(1. Beijing Institute of Electronic System Engineering, Beijing 100039, China;

2. PLA University of Science and Technology, Nanjing 210007, China;

3. State Key Laboratory of Software Engineering, Wuhan University, Wuhan 430072, China)

[Abstract] The distribution function is an important tool for the study of the stochastic variances. The normal distribution is very popular in the nature and human society. The idea of membership functions is the foundation of the fuzzy sets theory. While the fuzzy theory is widely used, the completely certain membership function which has no any fuzziness at all has been the bottleneck of the applications of this theory. Cloud models are effective tools in transforming between qualitative concepts and their quantitative expressions. It can represent the fuzziness and randomness and their relations of uncertain concepts. Also cloud models can show the concept granularity in multi-scale spaces by the digital characteristic Entropy (E_n). The normal cloud models not only broaden the formation conditions of the normal distribution but also make the normal membership function be the expectation of the random membership degree. In this paper, the universality of the normal cloud model is proved, which is more superior and easier, and can fit the fuzziness and gentleness in human cognizing process. It would be more applicable and universal in the representation of uncertain notions.

[Key words] normal distribution function; membership degree; normal cloud model; granularity; universality