View metadata, citation and similar papers at core.ac.uk
brought to you by CORE
provided by PolyU Institutional Repository

第 31 卷 第 24 期 计 算 机 工 程 2005 年 12 月
Vol.31 № 24 Computer Engineering December 2005
·博士论文· 文章编号：1000—3428(2005)24—0036—02 文献标识码：A 中图分类号：O236

熵及其在空间数据不确定性研究中的应用

史玉峰[1,2]，史文中[3]，靳奉祥[4]
(1. 山东理工大学建筑工程学院，淄博 255049；2. 武汉大学地球空间环境与大地测量教育部重点实验室，武汉 430079；
3. 香港理工大学土地测量与地理资讯系地球资讯科技研究中心，九龙；4. 山东科技大学校长办公室，青岛 266510)

摘 要：总结了熵的产生、发展、特性及其应用，讨论了熵与不确定性的关系；针对空间数据不确定性问题，总结了基于熵的空间数据不确定性研究成果，提出了应用混合熵作为统一测度来度量空间数据不确定性的设想。

关键词：熵；不确定性；空间数据；信息熵；模糊熵；混合熵

Entropy and Its State of Arts on Research of Spatial Data Uncertainty

SHI Yufeng[1,2], SHI Wenzhong[3], JIN Fengxiang[4]
(1. School of Architecture Engineering, Shandong University of Technology, Zibo 255049; 2. Key Laboratory of Geospace Environment and Geodesy, Ministry of Education, Wuhan University, Wuhan 430079; 3. Advanced Research Center for Spatial Information Technology, Department of Land Surveying and Geo-Informatics, HongKong Polytechnic University, Hunghom, Kowloon, HongKong;
4. President Office, Shandong University of Science & Technology, Qingdao 266510)

【Abstract】 In this paper, the definition of entropy and its characteristics as well as its developments are summed up firstly, and then the relationship of entropy and uncertainty are discussed. Aiming at the problem of spatial data uncertainty, this paper sums up the proposed output of spatial data uncertainty based on entropy, and presents that hybrid entropy can be used as the uniform measure of spatial data uncertainty.

【Key words】 Entropy; Uncertainty; Spatial data; Information entropy; Fuzzy entropy; Hybrid entropy

1 熵的提出

熵的概念是由德国物理学家 Clausius 于 1865 年提出的。最初熵是建立在热力学第二定律基础之上的，用以描述自发过程不可逆性的状态函数，但这一定义仅能描述宏观过程的不可逆性，却不能反映体系内部的结构变化特征。为了解释不可逆过程的微观机理，Boltzmann 给出了如下形式的熵函数：S=klogW，这里的 k 为 Boltzmann 常数，W 为与某一宏观状态所对应的微观状态数，即系统处于某一状态的概率。Boltzmann 熵函数把宏观量 S 与微观状态数 W 联系起来，在宏观与微观之间假设了一座桥梁，既说明了微观状态数的物理意义，又给出了熵函数的统计解释(微观意义)。

2 信息熵及其应用

现在人们应用和研究的熵一般指信息熵或由信息熵演化生成的其它熵。信息熵的概念是由信息论的创始人 Shannon 于 1948 年提出的。信息理论是应用统计方法的延伸，它规定信息量等于消除的不确定性的数量；若所有不确定性都消除了，则信息量为最大。对于取值离散的样本空间（信源）

$$[X \bullet P]:\{X:a_1,a_2,\cdots,a_r \quad P(X):p_1,p_2,\cdots,p_r\}$$

其中 p_i 为事件 a_i 出现的概率，则事件 a_i 的所含有的信息量(自信息)用 $I(a_i)$表示，即

$$I(a_i)=\log\frac{1}{p_i}=-\log p_i \tag{1}$$

自信息 $I(a_i)$含有两种意义：在事件 a_i 发生以前，表示其不确定性；事件 a_i 发生以后，则表示其所提供的信息量。由于自信息 $I(a_i)$是一个随所发生消息变化的随机变量，不宜用作整个样本空间信息的度量，因此 Shannon 在信息论中定义自信息的数学期望为信源的信息熵，即

$$H(X)=E[-\log p_i]=-\sum_{i=1}^{n}p_i\log p_i \tag{2}$$

信息熵具有下列主要性质：

（1）对称性

$H(p_1,p_2,\cdots,p_r)=H(p_2,p_1,\cdots,p_r)=H(p_r,p_2,\cdots,p_1)$；

（2）非负性

$H(p_1,p_2,\cdots,p_r)=-\sum_{i=1}^{r}p_i\log p_i\geq 0$；

（3）确定性，若信源 X 的概率空间中的任一概率分量，$p_i=1$，则信源的信息熵一定等于 0，即 H(X)=0；

（4）可加性，两个统计独立信源 X 和 Y,其信源空间分别为

$$[X \bullet P]:\{X:a_1,a_2,\cdots,a_r \quad P(X):p_1,p_2,\cdots,p_r\}$$

其中

$$0\leq p_i\leq 1,(i=1,2,\cdots,r),\sum_{i=1}^{r}p_i=1;$$

$$[Y \bullet P]:\{Y:b_1,b_2,\cdots,b_s \quad P(Y):q_1,q_2,\cdots,q_s\}$$

其中：$0\leq q_j\leq 1$，$(j=1,2,\cdots,s)$，$\sum_{j=1}^{s}q_j=1$。

若信源 X 和 Y 同时发出各自的符号，则构成联合信源(XY)；联合信源(XY)的概率分布 P(XY)等于：P(XY)=P(X)P(Y)，则联合信源(XY)的信息熵

$$H(XY)=-\sum_{i=1}^{r}\sum_{j=1}^{s}P(a_ib_j)\log p(a_ib_j)=H(X)+H(Y) \tag{3}$$

基金项目：山东省基础地理信息与数字化技术重点实验室开放基金资助项目(SD2003-10)；香港特区项目“ASD in Mitigation of Urban Hazards: Global Positioning System and Geographic Information System”(1.34.37.A222)

作者简介：史玉峰（1965－），男，博士、教授，主研方向：空间数据不确定性和空间数据分析等；史文中、靳奉祥，博士、教授

收稿日期：2004-12-27 **E-mail**：shyf@sdut.edu.cn

?1994-2017 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

(5)极值性，对任一信源空间，其熵满足：$H(p_1,p_2,\cdots,p_r)$；当信源是等概场时，其熵最大，为 $\log r$。

信息熵已在许多领域得到应用。一些学者用熵作为水文系统复杂性的统计测度，从系统可达状态的宏观概率层次和状态内部微观层次上的变化共同描述水文系统复杂性和不确定性；还有学者应用信息熵解决了水文系统概率分布推导和参数估计问题、水文水质站网布设评估问题、水文水质站网布设评估问题等；基于信息熵，一些研究人员对降雨时空分布的不均匀性进行分析，由熵的空间分布状况，研究了信息有向传输问题，初步揭示了降水在空间分布的结构特征。

3 模糊熵及其应用

1965 年，Zadeh 第一次提出了“模糊集合”概念。此后，众多学者对模糊理论及其相关问题进行了研究。受 Shannon 信息熵的启发，Zadeh 最先提出度量模糊事件不确定性的设想，用模糊熵度量模糊子集的模糊不确定性[4]。

设集合 $U=\{x_1,x_2,\cdots,x_n\}$ 是有限的，A 为 U 上的一个模糊子集，即 $\mu_A:U\to[0,1]$。$\mu_A(x_i)$ 是 U 中元素 x_i 隶属于 A 的隶属度。则模糊熵定义为：模糊熵 H_f 是将 U 的幂集 2^U 映射到非负实数空间。即：$H_f:2^U\to[0,\infty)$，并且，对于 U 上任意两个模糊子集 A 和 B，满足下列 4 条公理:

（1）若 A 是一明晰集,对任意 $x_j\in U,\quad j=1,2,\cdots,n$，有 $\mu_A(x_i)=\{0,1\}$，则 $H_f(A)=0$。

(2)对任意 $x_j\in U,\quad j=1,2,\cdots,n$，当 $\mu_A(x_i)=0.5$ 时，$H_f(A)$ 取得最大值。

(3)若 B 是 A 的任一较清晰形式，即：任意 $x_j\in U$，如果有 $\mu_A(x_i)\le 0.5$，则有 $\mu_B(x_i)\le\mu_A(x_i)$；若 $\mu_A(x_i)\ge 0.5$，则有 $\mu_B(x_i)\ge\mu_A(x_i)$，那么 $H_f(B)\le H_f(A)$。

(4)如 A^C 为模糊子集 A 的补集,则有 $H_f(A)=H_f(A^C)$。

许多研究人员根据所研究的领域和问题，建立了不同的模糊熵函数[1~3]。Hall 从几何意义上讨论了熵与不确定性和信息的关系，提出了总体熵(Ensemble Entropy)和总体体积(Ensemble volume)概念，讨论了熵、不确定性和信息的几何特性。

4 熵与不确定性的关系

前苏联学者诺维茨基在其论著《测量结果误差估计》中提出了应用信息熵作为测度，确定了随机变量的熵不确定度区间。根据信息论，信息量是熵的差值 $I=H(X)-H(X/x_0)$。对于一维观测值，在测量之前，根据已有的知识和经验，知道其取值范围[a,b]；观测 x_0 之后，由于观测值 x_0 不可避免地带有误差 $\pm\Delta$，因此只能肯定其测量值是落在区间 $[x_0-\Delta,x_0+\Delta]$ 内。从信息论的角度理解，测量的意义就是使不确定度区间长度由(b-a)缩短到 2Δ。对于一维观测值，在测量之前，根据已有的知识和经验，知道其取值范围[a,b]；观测 x_0 之后，由于观测值 x_0 不可避免地带有误差 $\pm\Delta$，因此只能肯定其测量值是落在宽度范围为 $d=2\Delta$ 的不确定区间内的某一点上，即在区间 $[x_0-\Delta,x_0+\Delta]$ 内。从信息论的角度理解，测量的意义就是使不确定度区间长度由 $(b-a)$ 缩短到 2Δ。假设测量以后概率密度分布函数为 $p(x)=1/2\Delta$，则观测以后测量结果的熵为

$$H(X/x_0)=-\int_{x_0-\Delta}^{x_0+\Delta}\frac{1}{2\Delta}\ln\frac{1}{2\Delta}dx=\ln 2\Delta$$

文献《测量结果误差估计》将 $\Delta_S=\frac{1}{2}d=\frac{1}{2}e^{H(X/x_0)}$ 称为“误差熵值”，则由误差熵值所确定的不确定度区间为“熵意义上的不确定度区间”。对于一维 Gaussian 分布，若其概率分布函数为 $f(x)\sim N(\mu,\sigma^2)$，其中 μ 为观测值的数学期望，σ^2 为方差。由信息熵的定义，可以求得 $H(X/x_0)=\ln\sqrt{2\pi e\sigma^2}$，误差熵为 $\Delta_s=2.066\sigma$，熵不确定度区间为 $[\mu-2.066\sigma,\mu+2.066\sigma]$，点位落在该区间的概率为 $P(\mu-2.066\sigma<x<\mu+2.066\sigma)=0.961$。在统计分析和误差处理领域，常用 2 倍中误差或 3 倍中误差作为不确定度区间，点位落入这些区间的概率分别为 95.4%和 99.7%；而基于熵的不确定度区间介于 2 倍和 3 倍中误差之间，它是应用观测误差分布求出的一个比较客观的指标。基于这一思想，史玉峰等提出基于信息熵的粗差识别方法[5]。

5 熵在空间数据不确定性中的应用与展望

熵在其诞生的 100 多年里，从最初用于定量地阐明热力学第二定律，发展成为现代自然科学和工程技术领域的有力数据分析工具。Shannon 所定义的信息熵是信息量的测度，由于其物理意义清晰，使用方便，得到了广泛的应用。半个多世纪来，随着信息科学的发展，信息熵已成为一种具有普遍意义的数据分析工具，并在多个学科中取得了丰硕的研究成果，人们开始采用崭新的信息熵方法来研究高级事物的复杂行为。在 GIS 领域，郭大志提出了整体 GIS 数据质量指标的概念，用条件信息熵来描述数据集的整体质量，研究了基于误差熵不确定带的数据质量评价模型。基于信息理论中信息熵的概念，范爱民应用线元端点边缘概率分布函数的误差熵作为确定“ε–带”带宽的尺度，建立了误差熵带(H-带)模型[7]。李大军、龚健雅等在范爱民的基础上，对“H-带”进行了改进，建立了平均熵带模型和熵意义下的面元不确定性模型“ε_σ 带”模型[6]。

由于空间数据不确定性的内涵十分丰富，表现形式多种多样，因此在空间数据不确定性研究问题上，需要有一个统一的测度来度量其不确定性。现有的许多理论方法，如信息论、模糊数学、粗糙集、证据理论、混沌理论等都用到熵作为测度，而且在许多领域中熵已经作为评价相关不确定性的测度。鉴于此，可否用熵作为统一尺度，来评价空间数据的不确定性？具体可以从下几方面进行研究：(1) 空间数据的不确定性不仅具有随机性，还含有模糊性；而现有空间数据不确定性的模糊性研究多是定性研究，可否以熵为测度建立空间数据不确定性的模糊熵模型。(2) 空间数据不确定性是传统误差概念的延拓，可以将其看作一种广义的误差。即不确定性既包含随机误差也包含系统误差和粗差，还包含可度量和不可度量的误差，以及数值上和概念上的误差，所有这些不确定性可否用熵来统一度量。(3) 将空间数据位置的随机性和属性的模糊性统一考虑，以熵为测度，建立空间数据不确定性的混合熵（或总体熵）模型。(4) 空间数据的未确知性和灰性是其另外两种不确定性，目前对这两种不确定性的研究较少，能否用熵理论来度量空间数据的未确知性和灰性也是一个值得研究的问题。(5) 熵可以作为 GIS 中数据信息不确定性测度，那么可否应用熵来对 GIS 中数据质量进行控制等都是值得探究的问题。

(下转第 43 页)

?1994-2017 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

4.2 MWPSO 与 GA 的仿真结果分析

设控制对象的数学模型为

$$G(s)=\frac{60e^{-s}}{(s+2)(s+5)}=\frac{6e^{-s}}{(0.2s+1)(0.5s+1)}$$

采样时间为 20ms，输入信号为阶跃信号。

MWPSO 算法中初始化种群大小为 40，搜索空间为待优化参数 k_p、k_i、k_d 所确定的 3 维空间。为了避免参数选取的范围过大而可能造成的寻优盲目性，可以根据经验选取一组参数范围，这样可以节约寻优的时间和计算量。参数 $k_p\in[0,2]$，$k_i\in[0,3]$，$k_d\in[0,1]$，取 w_1、w_2、w_3、w_4 的值分别为 0.999、0.001、100、2.0；种群的初始权重分别为 1.2、1.0、0.9、0.8。遗传算法的种群大小也设置为 40，两种算法的终止代数均为 200，交叉率 P_c=0.9，变异率 P_m=0.033。将 MWPSO 算法与 GA 算法各连续运行 20 次，两种方法 20 次优化计算过程中平均的 J_{best} 曲线如图 1 所示。各次运行的目标函数值如图 2 所示。

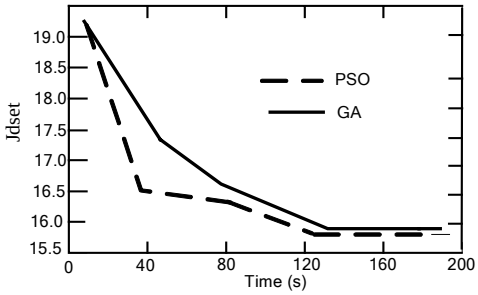

图 1　目标函数 J_{best} 的优化过程

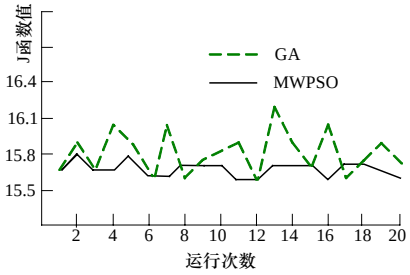

图 2　各次运行 J 函数值对比

经过 200 代后，GA 整定的结果为 k_p=0.986 8，k_i=0.526 4，k_d=0.335 6，目标函数 J=15.974 6。MWPSO 整定的结果为 k_p=0.712 5、k_i=0.482 6、k_d=0.299 7，目标函数 J=15.768 3。两种算法整定结果的动态响应比较如图 3 所示。通过两种优化算法对 PID 参数的寻优，系统具有更好的动态性能。同时由图 1~图 3 看出，准确寻找到最优 J 函数值对控制效果的改善有明显的作用；而在相同的迭代次数下文中的 MWPSO 算法比 GA 算法更能准确快速地找到最优值；且 MWPSO 算法能有效地保证单次运行结果。

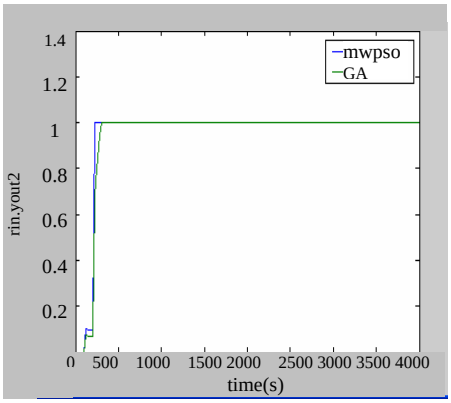

图3 单位阶跃响应曲线

5 结束语

本文提出了一种改进粒子群算法，即 MWPSO 算法，并用 4 个常用的测试函数进行测试，测试结果说明该算法是高效、实用的。利用 MWPSO 算法对 PID 参数进行寻优，将其结果与 GA 寻优的结果进行了比较。计算和仿真结果表明，在相同的迭代次数下，MWPSO 优化 PID 参数的指标好于 GA 算法，控制效果也有一定的提高。

参考文献

1 Eberhart R C, Shi Yuhui. Comparison Between Genetic Algorithms and Particle Swarm Optimization[C]. Annual Conference on Evolutionary Programming, San Diego, 1998

2 Kennedy J, Berthart R. Particle Swarm Optimization[C]. In: Proc. of IEEE Int. Conf. on Neural Networks, Perth, 1995: 1942-1948

3 Shi Yuhui, Eberhart R C. Fuzzy Adaptive Particle Swarm Optimization[C]. In: Proc. of IEEE Int. Conf. on Evolutionary Computation. Seoul, 2001: 101-106

4 Shi Y, Eberhart R C. A Modified Swarm Optimizer[C]. IEEE International Conference of Evolutionary Computation. Anchorage, Alaska: IEEE Press, 1998-05

5 Clerc M, Kennedy J. The Particle Swarm: Explosion, Stability, and Convergence in Multi-Dimension Complex Space[J]. IEEE Transactions on Evolutionary Computation, 2002, 16(1): 58-73

6 Eberhart R, Shi Y. Particle Swarm Optimization: Development, Applications and Resource[C]. IEEE Int. Conf. on Evolutionary Computation, 2001: 81-86

7 李　萌, 沈　炯. 基于自适应遗传算法的过热气温 PID 参数优化控制仿真研究[J]. 中国电机工程学报, 2002, 22(8): 145-149

(上接第 37 页)

参考文献

1 Bhandri D, Pal N R. Some New Information Measure for Fuzzy Sets [J]. Information Science, 1986, 67(3): 165-174

2 Hall J W. Universal Geometric Approach to Uncertainty, Entropy and Information [J]. Physical Review (A), 1999, 59(4): 2602-2615

3 Yager R R. Measure of Entropy and Fuzziness Related to Aggregation [J]. Information Science, 1995, 82(3-4): 147-166

4 Zadeh L A. Probability Measures of Fuzzy Events [J]. Journal of Mathematics Analysis and Application, 1968, 23(10): 421-427

5 史玉峰, 靳奉祥, 王　健. 基于信息熵的测量粗差识别方法[J]. 测绘通报, 2003, 33(3):9-11

6 李大军, 龚健雅, 谢刚生等. GIS 面元的误差熵模型[J]. 测绘学报, 2003, 32(1): 31-35

7 范爱民, 郭大志. 误差熵不确定带模型[J]. 测绘学报, 2001, 30(1): 48-53

?1994-2017 China Academic Journal Electronic Publishing House. All rights reserved.　http://www.cnki.net