用统计物理的方法计算信源熵率

陈双平[①] 郑浩然[②] 马 猛[②③] 张振亚[①] 王煦法[②]
[①](中国科学技术大学电子工程与信息科学系 合肥 230027)
[②](中国科学技术大学计算机科学与技术系 合肥 230027)
[③](安徽大学计算机科学与技术学院 合肥 230039)

摘 要：从数学模型的角度来说，信源和随机过程有着一一对应的关系。从混沌的角度来看，随机过程的多重分形谱是动力系统的重要特征，熵率只是多重分形维中特殊的一维，即信息维。该文指出了如何用统计物理的方法计算随机过程的多重分形维，以二态隐马尔可夫信源作为例子，该文计算了其熵率。计算结果和理论结果的比较表明，用统计物理的方法计算隐马尔可夫过程熵率具有实用价值。这一方法可以推广到一般信源熵率的数值计算。

关键词：信源；熵率；多重分形谱；隐马尔可夫过程
中图分类号：TN911.2 **文献标识码**：A **文章编号**：1009-5896(2007)01-0129-04

Computing the Entropy Rate of Information Source with Methods of Statistical Physics

Chen Shuang-ping[①] Zheng Hao-ran[②] Ma Meng[②③] Zhang Zhen-ya[①] Wang Xu-fa[②]
[①](*Department of Electronic Engineering and Information Science*, *University of Science and Technology of China* (*USTC*), *Hefei* 230027, *China*)
[②](*Department of Computer Science and Technology*, *USTC*, *Hefei* 230027, *China*)
[③](*School of Computer Science and Technology*, *Anhui University*, *Hefei* 230039, *China*)

Abstract: From the mathematical point of view, information sources can be 1-to-1 mapped to stochastic processes. Known from the theory of chaos, multi-fractal of stochastic process is a key characteristic of its dynamics, of which entropy rate is a special fractal dimension named information dimension. The paper introduces methods of statistical physics to compute the multi-fractal of stochastic process so that the entropy rate of source can be obtained at once. Take binary hidden Markov processes as example, the paper demonstrate how this approach works. The results shows that the methods is applicable to numerically approximate the entropy rate of binary hidden Markov processes (BHMPs) in practical applications, and it can be applied in more generalized kinds of information sources.

Key words: Information souce; Entropy rate; Multi-fractal; Hidden Markov processes

1 引言

信源输出信号在数学中可以用随机过程加以描述，因此，可以说信源的建模在某种程度上也就是用恰当的随机过程来描述信号[1]。然而，一般的随机过程理论并不涉及和讨论信号中所携带的信息，而信息论所关心的中心内容则是信号中携带的信息。在信息论中信源输出信号所携带信息的效率使用熵率或冗余度来表示的，因而各种信源的熵率的计算方法就成为信源编码中的主要研究内容[1]。

离散无记忆信源的熵率是模型概率参数的一个解析表达式；离散马尔可夫信源的熵率也可以由并不复杂的方法计算得到[1]。但是对于一般的信源，如隐马尔可夫信源[2,3]，其熵率目前没有找到解析表达式[4]。我们在以前的相关工作中，对一类特殊的二态隐马尔可夫信源的熵率，证明了其一个收敛的上下界[5]，并且给出了一个理论上可以逼近到任意精度的数值算法[6](理论上熵率的误差可以计算到低于任意小的正值)。但是，这并没有解决一般的隐马尔可夫过程熵率的计算问题。对于更一般的随机过程所建模的信源，这一问题同样存在。

从复杂性科学的角度来说，熵率只是非线性系统多重分形谱中特殊的一维，即信息维[7]。因此熵率的计算完全可以用统计物理的方法加以解决。多重分形可以分为规则分形和不规则分形。规则多重分形可以用解析方法或统计物理的方法得到它们的多重分形谱，不规则多重分形谱只能用统计物理的方法得出[7]。在得到其多重分形谱以后，熵率就可以从多重分形谱中计算得到。

2005-05-23 收到，2005-09-22 改回
中国科技大学高水平大学建设重点项目和中国科学技术大学青年基金资助课题

在本文中，我们引入统计物理中的方法，用以计算信源的熵率。我们将信源的随机模型和非线性动力学中对象之间建立联系，其中离散无记忆信源、马尔可夫信源和康托集之间具有对应关系。我们还将演示如何用解析的方法计算离散无记忆信源的熵率；如何用统计物理的方法计算隐马尔可夫信源的熵率。并且这种统计物理的方法，完全可以推广到一般信源熵率的计算。

2 定义和记号

2.1 不均匀康托集的多重分形谱和熵率

在下面我们简单介绍一下康托集和多重分形谱的概念，关于更多有关康托集和多重分形谱的介绍请参见文献[7]。

一种质量分布不均匀的简单康托二分集可以由如下方式生成：初始只有一条线段，每操作一次，将原有线段三等分并舍去中间1/3段后，余下两段的质量分布概率分别为 P 和 $1-P$,即生成元是 $P/0/(1-P)$ 分布；接着再在两个1/3线段内分别用生成元作一次类似的操作，这样4个线段的质量分布概率有 3 种，概率最小的线段对应于 P^2，因为在两次操作中它的概率都取 P；中间两段的概率对应于 $P(1-P)$，这是一次取 P，一次取 $1-P$ 的结果；概率最大的线段对应于 $(1-P)^2$，显然它是两次取 $1-P$ 后得到的。这样操作 k 次以后,总线段数有 2^k 个，每个线段的尺寸 $\varepsilon=(1/3)^k$，其质量分布概率分别为 $P_i(\varepsilon)=P^m(1-P)^{k-m}, m=0,1,\cdots,k$。具有相同概率 P_i 的线段数分别为：$N(P_i)=k!/(m!(k-m)!)$。各个线段尺寸 ε 的 P_i 和 $N(P_i)$ 形成一个集(整个集延伸到 $k=\infty$)。

可以把全部概率分布 $P_i(\varepsilon)$ 组成的集划分为一系列子集，即按 $P_i(\varepsilon)$ 的大小划分为满足下面的幂函数的子集 $P_i(\varepsilon)\propto\varepsilon^{\alpha}$，这里的 α 是奇异指数。它是反映分形上各个小线段的奇异程度的一个量，所以 α 的数值必然是与所在的子集有关。若在分形上的测度量(如前述的质量)是均匀的，则 α 值必然只有一个值。若不均匀，可以用 α 值的大小区分为许多小子集。将子集内的线段数或单元数 $N(\varepsilon)$ 和 ε 的关系式定义为[7]：$N(\varepsilon)\propto\varepsilon^{-f(\alpha)}, \varepsilon\to 0$。$f(\alpha)$ 的物理意义是表示相同 α 值的子集的分形维数。一般将 $f(\alpha)$ 称为多重分形谱。康托集的多重分形谱 $f(\alpha)$ 可以解析求解。

统计物理给出了规则和不规则多重分形谱的计算方法。首先定一个配分函数 $\chi_q(\varepsilon)$，对概率 $P(\varepsilon)$ 用 q 次方进行加权求和，其数学表达式为[7]：$\chi_q(\varepsilon)\equiv\Sigma P_i(\varepsilon)^q=\varepsilon^{\tau(q)}$。如果上式后面的等式成立，即配分函数和 ε 有幂函数关系，则可以从 $\ln\chi_q(\varepsilon)\sim\ln\varepsilon$ 曲线的斜率得到[7] $\tau(q)=\left[\ln\chi_q(\varepsilon)\right]\Big/(\ln\varepsilon),(\varepsilon\to 0)$，一般把 $\tau(q)$ 成为质量指数。

当 $q\gg 1$ 时，在 $\Sigma P_i(\varepsilon)^q$ 的求和中大概率子集将起到主要作用；$q\ll -1$ 时，$\Sigma P_i(\varepsilon)^q$ 求和中小概率子集将起到主要作用。所以通过加权处理，可以对一个分形集内部的结构进行精细的研究。

广义分形维数 D_q 定义为[7]：$D(q)=\left[\tau(q)\right]\Big/(q-1)=\left[\ln\chi_q(\varepsilon)\right]\Big/\left[(q-1)\ln\varepsilon\right],(\varepsilon\to 0)$。它是随不同的 q 值而有不同意义的分形维数。$q=0$ 时，D_0 是普通的豪斯道夫维数。当 $q=1$ 时，可推导出：$D_1=\left(\Sigma P_i\ln P_i\right)\Big/(\ln\varepsilon),(\varepsilon\to 0)$。这里 $\Sigma P_i\ln P_i$ 相当于系统的负熵，也就是系统的信息熵，所以 D_1 是信息维数，也就是熵率。

2.2 二态隐马尔可夫过程和带标签的康托集

令 $X=\{X_k\}_{k\ge 1}$ 为二元字符集上的一阶静态马尔可夫过程，其转移概率 $P=\pi_{ab}$ 满足 $\pi_{ab}=P_X(X_k=b\mid X_{k-1}=a), a,b\in\{0,1\}$。另外有一个伯努利(Bernoulli)噪声过程(二态独立同分布) $E=\{E_k\}_{k\ge 1}$，独立于 X，满足 $P(E_i=1)=\varepsilon$。因此，可以定义随机过程 $Z=\{Z_k\}_{k\ge 1}$，有 $Z_k=X_k\oplus E_k, k\ge 1$，这里 $\oplus$ 指模 2 加法(“异或”)。可以将 Z 视为具有噪声 E 的二元对称信道的输出，其输入为 X。Z 完全由参数 π_{01}，π_{10} 和 ε 决定。随机过程 Z 是一种最简单的隐马氏过程，简称为二态隐马尔可夫过程[8](Binary Hidden Markov Process，BHMP)。通常情况下我们认为噪声 ε 较小，只考虑 $0<\varepsilon<\pi_{01}$，$\pi_{10}<1/2$ 的情形。本文中，我们给出的例子为 $\pi_{01}=\pi_{10}=\pi$ 时的情形。

我们业已证明，含 M 个隐状态 N 种输出值的隐马尔可夫过程和一个带 M 维标签的 N 分不均匀康托集一一对应[8]，其中 M 维标签为输出某个序列时最后所处隐状态的比例。以二元隐马尔可夫信源为例，它可以相当于一个带二维标签的不均匀二分康托集。这种康托集经过 k 次分裂后，产生 2^k 个输出 $Z_1^k\equiv z_1\cdots z_k$，集合中的每个元素 $Z_1^k=z_1^k$ 具有标签 $(P(X_n=0\mid Z_1^n),1-P(X_n=0\mid Z_1^n))$ 和概率 $P(Z_1^{n+1})$。标签和概率的计算由以下的公式得出。

定义 $g_0(x)$ 和 $g_1(x)$ 为[6]：$g_0(x)=(1-2\pi)(1-2\varepsilon)x+\cdot\pi(1-\varepsilon)+(1-\pi)\varepsilon, g_1(x)=1-g_0(x)$。因此[6]

$$P(Z_{n+1}=0\mid Z_1^n)=g_0(P(X_n=0\mid Z_1^n)) \tag{1}$$

$$P(Z_{n+1}=1\mid Z_1^n)=g_1(P(X_n=0\mid Z_1^n)) \tag{2}$$

因此有[6]

$$P(Z_1^{n+1})=P(Z_{n+1}\mid Z_1^n)P(Z_1^n)=g_{z_{n+1}}(P(X_n=0\mid Z_1^n))P(Z_1^n) \tag{3}$$

定义 $f_0(x)$ 和 $f_1(x)$ 为[6] $f_0(x)=\left[(1-\varepsilon)(x(1-2\pi)+\pi)\right]\Big/\left[(1-2\pi)(1-2\varepsilon)x+\pi(1-\varepsilon)+(1-\pi)\varepsilon\right], f_1(x)=\left[\varepsilon(x(1-2\pi)+\pi)\right]/\left[-(1-2\pi)(1-2\varepsilon)x+\pi\varepsilon+(1-\pi)(1-\varepsilon)\right]$。因此有[6]

$$P(X_n=0\mid Z_1^{n-1}0)=f_0(P(X_{n-1}=0\mid Z_1^{n-1})) \tag{4}$$

$$P(X_n=0\mid Z_1^{n-1}1)=f_1(P(X_{n-1}=0\mid Z_1^{n-1})) \tag{5}$$

3 计算熵率的统计物理方法

3.1 熵率的混沌学意义及其计算

引理 1[9,10] 多重分形谱的图像 $y=f(\alpha)$ 与直线 $y=\alpha$，相切于点 (D_1,D_2)，即 $D_1=f(D_1)$。

在文献[9,10]中提到判断多重分形谱是否正确的条件中也包含这两点。这个定理为我们求解熵率提供了手段，熵率

就是多重分形谱 $y=f(\alpha)$ 与直线 $y=\alpha$ 的切点(交点)处 α 的值。若用解析的方法求熵率，只要解 $\alpha=f(\alpha)$ 这个方程即可；如果用统计的方法，则在图像上找到这个切点(交点)就可以了。

另外，混沌系统的空间分布维数对于熵率的计算而言也是必要的，我们已经知道有如下引理：

引理 2 [9,10]　多重分形谱的图像 $y=f(\alpha)$ 的最大值为 $f(\alpha)_{\max}$，它等于系统的豪斯道夫维数。即

$$D_0=f(\alpha)_{\max} \tag{6}$$

在此基础上，我们可以将信源变换到对应的非线性系统，从而计算其分形谱，并通过其分形谱计算信源的熵率。并且我们还证明，这种变换使得信源熵率和对应的非线性系统的空间维无关。如果有 M 种字符，则我们有

$$h=\left(D_1/D_0\right)\log_2(M) \tag{7}$$

这是因为信源的熵率反映的是全部可能信号序列所含的概率的分布情况，对每一输出信号序号都认为其对熵率的贡献是等同的，对于概率为零的输出信号序号是不考虑的。故有 $h=-\left[\ln(\chi_1)\right]/\left[\ln(N_{p>0})\right]\cdot\log_2(M)$，$D_1=-\left[\ln(\chi_1)\right]/\left[\ln(N_\varepsilon)\right]$，$D_0=-\left[\ln(N_{p>0})\right]/\left[\ln(N_\varepsilon)\right]$，其中 $N_{p>0}$ 为输出序列中概率大于零的数目，N_ε 为总的输出序列数，熵率的单位是比特，D_1/D_0 是以 M 为底的对数，故 $\log_2(M)$ 为归一化参数熵率的系数。由上易见式(9)成立。由此我们有算法 1，可以它来计算一般信源的熵率：

算法 1　计算信源的多重分形谱

(1) 构建信源到多重分形之间的映射。类似于康托集的构造过程，信源的输出字符集如果有 M 种的话，这个集合的每个元素等分为 M 份，分别计算 M 份的概率分布情况。这样，一个信源就变换为一个多重分形。

(2) 用解析或者数值计算的方法计算多重分形谱，用标度不变性检验得到的多重分形谱的有效性。统计物理的计算方法参见文献[11]。

(3) 根据引理 1 获得熵率和几何分形维。

(4) 应用式(7)修正熵率。

3.2　解析法求离散无记忆信源熵率

根据算法 1，我们可以将一个简单的二元离散无记忆信源用二项分布加以描述，而二项分布可以视为一个规则的多重分形，即康托集上的概率分布。康托集的多重分形谱 $f(\alpha)$ 可以解析求解，其计算过程可以参见文献[7]。α 和 $f(\alpha)$ 表示成参变量 ξ 的函数，其中 $\xi\in[0,1]$，且 $\alpha=-\xi\log_3 P+(1-\xi)\log_3(1-P)$，$f(\alpha)=-\xi\log_3\xi-(1-\xi)\log_3(1-\xi)$。我们将其分形谱画在图 1 中，图中直线为 $y=\alpha$，曲线为 $y=f(\alpha)$，直线和多重分形谱的交点(也是切点)为信源的熵率。从图 1 不难看出 $\alpha=f(\alpha)$ 的解为 $\xi=P$。$\xi=P$ 时，$\alpha|_{\xi=P}=f(\alpha)|_{\xi=P}=-P\log_3 P-(1-P)\log_3(1-P)$。$\xi=0.5$ 时，$f(\alpha)$ 取到 $\xi\in[0,1]$ 中的唯一极大值 $f(\alpha)|_{\xi=0.5}=\log_3 2$。故由式(7)知熵率为

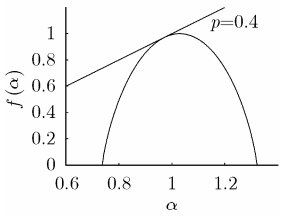

图 1　离散无记忆信源(二项分布)的多重分形谱

$$h=\frac{D_1}{D_0}=\frac{\alpha|_{\xi=P}}{f(\alpha)|_{\xi=0.5}}=-P\log_2 P-(1-P)\log_2(1-P) \tag{8}$$

这个式子也正是离散无记忆信源的熵率公式[1]。

3.3　数值计算法求隐马尔可夫信源的熵率

如前所述，隐马尔可夫信源可以等价为一个类康托集上的多重分形，按照算法 1 产生全部概率后，可以用数值的方法计算其分形谱。产生概率的算法伪代码参见算法 2，如要产生 14 次分裂后的具有 2^{14} 个元素的类康托集，调用方式为 generator(1,0.5,1,14,1)。

算法 2　产生 BHMP 的概率分布

```
function generator( p , p0 , k , kmax , i )
    if  k = kmax then
    printf  i , p
    return
    end if
    generator( p·g0(r0) , f0(p0) , k+1 , kmax , 2i−1 )
  generator( p·g1(r0) , f1(p0) , k+1 , kmax , 2i )
end function
```

取 $\pi=0.3$，$\varepsilon=0.2$，用算法 1 产生二元隐马尔可夫信源的概率分布以后，计算其配分函数 $\chi_q(\varepsilon)$，由图 2 中可以看出明显的线性关系，即有标度不变量存在，故其多重分形谱存在。用统计物理的方法[11]求得其多重分形谱并画在图 3 中。在图 3 中我们可以看到熵率为直线 $y=\alpha$ 和多重分形谱 $y=f(\alpha)$ 的交点处的值，值为 0.983997。在文献[6]中，我们提出了一种方法精确估计二元隐马尔可夫信源的熵率和偏差范围，估计的值为 0.983888 ± 0.0000005。可见用统计物理方法得到的熵率和理论值的偏差是很小的，约 0.00011。我们用大量的 (π,ε) 参数组合加以测试，都表明用统计物理方法得到的结果和理论值的误差在 0.001 以内，因此这种方法完全可以用于实际的应用。

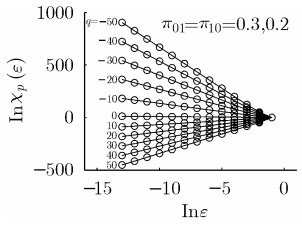

图 2　二元隐马尔可夫信源的 $\ln\chi_q(\varepsilon)$ 随 $\ln\varepsilon$ 的变化

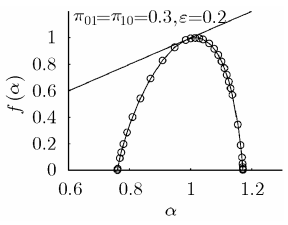

图 3　二元隐马尔可夫信源的多重分形谱

4 结束语

信源和随机过程有着对应的关系，随机过程的多重分形谱体现了其混沌动力学特征。在本文中，我们探讨了如何用统计物理的方法计算信源的熵率，熵率的计算和随机过程的多重分形谱的求解之间存在着相关性，可以通过解析或者数值的方法求得信源的多重分形谱，再求解熵率。实际上，对离散无记忆信源和马尔可夫信源，可以通过解析的方法求解其熵率；对于更加一般的信源，只要保证其多重分形的存在性，熵率可以用数值计算的方法解决。我们以离散无记忆信源和二态隐马尔可夫信源作为例子，显示了这种方法的有效性。本文的方法有助于理解信源的动力学行为和计算复杂信源的熵率。

参 考 文 献

[1] 朱雪龙. 应用信息论基础. 北京: 清华大学出版社, Mar. 2000: 74–108.

[2] Ephraim Y and Merhav N. Hidden Markov processes. *IEEE Trans. Inform. Theory*, 2002, 48(6): 1518–1569.

[3] Rabiner L R. A tutorial on hidden Markov models and selected applications in speech recognition. *Proce. IEEE*, 1989, 77(2): 257–286.

[4] Jacquet P, Seroussi G, and Szpankowski W. On the entropy of a hidden Markov process. In: Proceeding of the Data Compression Conference, Snowbird, Oct. 2004: 362–371.

[5] Chen S, Zheng H, Liu H, and Wang X. Estimators for the entropy rate of binary hidden Markov processes. http://prep.istic.ac.cn/docs/1111587263574.html, 2005.

[6] 陈双平, 郑浩然, 童庆, 王煦法. In 模糊逻辑与计算智能研究进展(2005 年论文集). 广东深圳, Apr. 2005. 合肥: 中国科学技术大学出版社: 932–938.

[7] 孙霞, 吴自勤, 黄昀. 分形原理及应用. 合肥: 中国科学技术大学出版社, 2003: 53–88.

[8] 陈双平, 郑浩然, 王煦法, 黎志升. 隐马尔可夫模型的混沌动力学分析. submited to Science in China Series E: Technological Sciences, http://prep.istic.ac.cn/docs/1111587035327.html,2005.

[9] 周炜星, 王延杰, 于遵宏. 多重分形奇异谱的几何特性: i.经典 renyi 定义法. 华东理工大学学报, 2000, 26(4): 385–389.
Zhou Wei-xing, Wang Yan-jie, Yu Zun-hong. Geometrical characteristics of singularity spectra of multifractals: Ⅰ. classical renyi definition. *Journal of East China University of Science and Technology*, 2000, 26(4): 385–389.

[10] 周炜星, 王延杰, 于遵宏. 多重分形奇异谱的几何特性: ii.配分函数法. 华东理工大学学报, 2000, 26(4): 390–395.
Zhou Wei-xing, Wang Yan-jie, Yu Zun-hong. Geometrical characteristics of singularity spectra of multifractals: Ⅱ. partition function definition. *Journal of East China University of Science and Technology*, 2000, 26(4): 390–395.

[11] Mach J, Mas F, and Sagues F. Two representations in multifractal analysis. *Journal of Physics A: Mathematical and General*, 1995, 28(19): 5607–5622.

陈双平： 男，1976 年生，博士后，研究兴趣为数据挖掘、生物信息学和复杂性.

郑浩然： 男，1967 年生，副教授，研究兴趣和方向为计算智能、生物信息学、模式识别.

马 猛： 男，1978 年生，博士生，研究兴趣为智能计算和生物信息学.

张振亚： 男，1972 年生，博士后，主要研究领域为信息检索、数据挖掘.

王煦法： 男，1948 年生，博士生导师，主要研究领域为计算机网络、计算智能、信号处理和模式识别等.