一种基于编码的混沌密钥流生成方法

温 涛[1,2] 张 永[1] 郭 权[2] 李凤坤[3]
(东北大学软件中心 沈阳 110004)[1] (东软信息学院计算机科学技术系 大连 116023)[2]
(东软信息学院教育资源开发与实训中心 大连 116023)[3]

摘 要 从信息论的角度对混沌密钥流的产生过程进行了描述,提出并证明了利用编码方法对实值混沌序列进行离散化可以提高产生流密钥的效率。研究了编码位数与二进制序列的伪随机性和流密钥生成效率之间的关系。编码位数对二进制序列的伪随机性的影响不明显,总体标准差最大为0.0135,特别地当编码位数为18时,标准差有明显的收敛趋势;编码位数在6位至15之间时,生成效率总体较高并且相差不大,其中位数为12时,取得最高的效率。从理论和实验两方面分别对二域法和提出的抽样编码法进行了比较,结果显示,通过抽样编码来进行离散化,能够保留实值混沌序列较多的伪随机性信息,提高了混沌流生成效率,该方法对丰富流密钥生成方法是一次有益的尝试。

关键词 混沌,密钥流,抽样编码,流密钥生成
中图法分类号 TP309 **文献标识码** A

Method for Generating Key Stream Based on Encoding

WEN Tao[1,2] ZHANG Yong[1] GUO Quan[2] LI Feng-kun[3]
(Software Center, Northeastern University, Shenyang 110004, China)[1]
(Department of Computer Science and Technology, Neusoft Institute of Information, Dalian 116023, China)[2]
(Eduction Resources Department and Training Center, Neusoft Institute of Information, Dalian 116023, China)[3]

Abstract The paper formulated the generating procedure of key stream of chaos in the term of information theory, presentd and proved that generating efficiency of key stream can be improved through sampling-and- encoding method to process real number sequence of chaos. And the paper made research on the relationship between the amount of encoding bits and pseudorandomness and generating efficiency. The amount of encoding bits effects pseudorandomness of binary sequence a little, and the standard devia-tion of each sample is 0.0135 at most, specially, when 18, the standard deviations converge apparently. Generating efficiency keeps higher, when the amount of encoding bits falls into the interval of 6 through 15, and arrive to the highest point when 12. The paper compared the binary-field method and the proposed sampling-and- encoding method in terms of theoretic and experimental proof and showed that sampling-and- encoding method can reserve more pseudorandom information so that it can improve the generating efficiency of key stream. The sampling-and- encoding method is a beneficial tentative experience to enrich the generating ways of key stream.

Keywords Chaos, Key stream, Sample and encoding, Stream key generation

1 引言

混沌系统因其高度复杂的非线性动力学特性及对初始条件和参数依赖的高度敏感性而被广泛引入到安全领域,例如,保密通信系统[1]、软件水印[2]、数字图像加密[3]等。对非基于困难问题的密钥产生算法的探索和非线性混沌系统的研究催生了混沌密码学[5]。存在的文献较多地关注了算法的安全性,较少考虑算法的效率。本文深入分析了混沌动力学系统,认为通过从混沌迭代过程中获取更多的非线性信息,可以提高产生算法的效率。因此提出了基于抽样编码的混沌密钥产生算法。

通过混沌映射函数将一个密钥(或称种子)扩展为密钥流成为构造同步密钥流[7,8]的一种重要方法。在实值混沌伪随机序列基础上得到的二进制混沌伪随机序列是其中的一个类别。在这一类的混沌伪随机序列生成中,由实值混沌伪随机序列转化为二进制混沌伪随机序列直接关系到最终密钥流的质量,因而是一个关键步骤。从统计学的角度来描述某一混沌映射函数得到的混沌分布可用均匀分布,也即连接随机变量。由信息论可知,连续随机变量的自信息和信息熵都是无穷大的,即要表示连续随机变量需要无穷多位。已有文献[7,8]无一例外地只用2bits表示混沌分布情况,以下将称这类方法为二域法。事实上,文献的做法是先将混沌吸引子(分布域)分成了两个部分,用2bits(0或1)来表示某一点落在哪个区域里。文献间的区别只是这两个区域的分法不同。从信

到稿日期:2010-05-26 返修日期:2010-08-19 本文受国家自然科学基金(60803131)资助。
温 涛(1962—),男,博士,教授,博士生导师,主要研究方向为网络安全、知识组织,E-mail:wentao@neusoft.edu.cn;张 永(1981—),男,博士生,主要研究方向为网络安全、密码学,E-mail:zhangyong@neusoft.edu.cn(通信作者)。

息论的角度分析，这些方法没有本质的差别。只用 2bits 描述混沌分布，每次迭代只能产生 1bit 有用信息，导致其所体现的随机性信息在转化过程中被大量遗漏，因此，产生流密钥的效率比较低。当采用 n 轨道时，每次只能得到 $1/n$ bits 有用信息，导致产生相等长度的密钥需要更多次迭代，此时，效率降低更严重。

定义 1 称由实值混沌伪随机序列转化为二进制混沌伪随机序列的过程为离散化过程。

本文首先给出了离散化过程的信息论解释，研究了编码方案。最后提出了一种通过编码来实现离散化的方法并与二域法进行了比较。该方法在转化过程中较多地保留了混沌映射函数自身的随机性信息，从而能产生更多的密钥流位，最终显著地提高了密钥流产生器的效率。

本文第 2 节给出了离散化过程的信息论解释；第 3 节研究了编码方案并提出了一种离散化方法；第 4 节从理论和实验两个方面分析了离散化方法的性能和安全性，并与二域法做了比较；最后总结了全文所做的工作及以后要解决的问题。

2 离散化过程的信息论解释

从信息角度来看，离散过程是一个随机信息的传递过程。由数据通信可知，若将由某一区间$[a,b]$上的混沌分布看作是连续型随机变量的样本值时，离散化过程就是模拟信号转化为数字信号(A/D)的过程。A/D 转化的基本方法就是抽样和编码，并且要求转化过程保证信号不失真。

连续型随机变量的自信息量和信息熵按离散型随机变量相应概念推广后均为无穷大，也即蕴涵着无限信息。根据抽样定理，若有一频带限制在$(0,f_H)$赫内的时间连续信号 $m(t)$，如果以 $T\leqslant 1/2f_H$ 的间隔对它进行等间隔抽样，则 $m(t)$ 将被所得到的抽样值完全确定。也就是说，以 $T\leqslant 1/2f_H$ 为间隔得到的样本就包含了连续时间信号必要的信息。由此可见连续时间信号中包含着大量的冗余信息。有以下定理。

定理 1 设区间$[a,b]$上的连续型随机变量为 X。将区间$[a,b]$分成 k 个两两不相交的子区间 $A_1,A_2,\cdots,A_k$，且满足条件$(\bigcap_{i=1}^{k}A_i=\Phi)\cap(\bigcup_{i=1}^{k}A_i=[a,b])$，令正实数序列 $l_1,l_2,\cdots,l_k$ 表示区间长度，令事件序列 $X_1,X_2,\cdots,X_k$ 分别表示随机变量的值落在相应区间的事件。则事件序列的自信息量和信息熵能反映原连续型随机变量的必要信息。

证明：由信息论可知事件序列的自信息量为

$$I(X_i)=-\log(p(X_i))=-\log(l_i/(b-a)) \tag{1}$$

当 $\max(l_1,l_2,\cdots,l_k)\to 0$，有 $X_i\to X$，此时，又由式(1)得 $I(X_i)\to+\infty$。综上，X_i 可以代替 X。同理可证事件序列的信息熵为式(2)。

$$\begin{aligned}H(X)&=\sum_{i=1}^{k}p(X_i)I(X_i)\\&=-\sum_{i=1}^{k}(l_i\log(l_i/(b-a)))/(b-a)\end{aligned} \tag{2}$$

可以代替 X 的信息熵。

定理 1 表明事件序列包含着描述连续随机变量的近似信息。我们可以用这一事件序列来近似表达该随机变量所提供的信息。为了获取均匀分布的随机变量的整体信息，抽样的最佳方式是采用等间隔抽样。

定理 2 在定理 1 的基础上令 $l_1=l_2=\cdots=l_k=l$，则事件序列的自信息量和信息熵是分区数量的增函数。

证明：式(1)和式(2)变为式(3)和式(4)：

$$I(X_i)=-\log(p(X_i))=-\log(1/k) \tag{3}$$

$$\begin{aligned}H(X)&=\sum_{i=1}^{k}p(X_i)I(X_i)=-\sum_{i=1}^{k}(\log(1/k))/k\\&=-\log(1/k)\end{aligned} \tag{4}$$

易得 $I(X_i)$和 $H(X)$都是 k 的增函数。

定理 2 表明事件序列从其源信号中获取的信息量的多少与区间的数量有增函数关系。若要得到较多的信息可以通过增加区间数量来达到目的。

若混沌映射函数的混沌分布是均匀的，即有较好的统计特性，则离散化方法也应能保证离散化后的二进制序列是均匀分布的。

定义 2 某一状态集合的状态个数恰为 2^n，$(n=1,2,3,\cdots)$，若用 $n$bits 二进制数为其编码，则称这样的编码为全值编码。

定理 3 设区间$[a,b]$上的混沌吸引子为 X，其统计分布是均匀的。则通过均匀抽样和全值编码后得到的二进制序列是均匀分布的。

证明：按定理 1 中提到的方法将$[a,b]$分成 k 个两两不相交的子区间 $A_1,A_2,\cdots,A_k$，区间长度和事件表示不变。因为 X 是均匀分布的随机变量，所以有等式 $p(X_1)=p(X_2)=\cdots=p(X_k)$。对于相应的编码序列有等式 $p(s_1)=p(s_2)=\cdots=p(s_k)$成立，其中 $S_i=\{e_1,e_2\cdots e_n\}(i=1,2\cdots k)$表示对事件 X_i $(i=1,2\cdots k)$的编码。因为是全值编码，所以有：

$$k=2^n \tag{5}$$

由 $p(X_1)=p(X_2)=\cdots=p(X_k)$得：

$$p(X_i)=1/k(i=1,2,\cdots k) \tag{6}$$

另一方面，若序列 $s=\{e_1,e_2,\cdots,e_n\}$为独立实验序列，也即服从二项分布，则有：

$$p(s_i)=1/2^n(i=1,2,\cdots,k) \tag{7}$$

将式(5)代入式(6)后与式(7)相同。也即 $s=\{e_1,e_2,\cdots,e_n\}$是二项分布。命题得证。

定理 3 表明，若混沌吸引子是均匀分布的，则通过抽样编码后的二进制序列是服从二项分布的。因此，通过抽样编码后混沌动力学系统的伪随机性信息能够传递，体现为产生的流密钥的伪随机性。也即可以通过抽样编码来实现离散化。

3 编码方案及离散化方法

如何确定抽样频率(由全值编码可知，同时也就确定了编码的位数)是涉及到效率和安全性的重要问题。针对攻击方法来设计保护手段是密码学的常用方法。以下也将按这个思路来设计编码方案。有下面一般性(非正式的)定理。只给出说明没有证明。

定理 4 理想情况的一次一密加密体制下，只有唯密文攻击有意义。

根据密码分析者破译时已经具备的前提条件，通常人们将攻击类型分为下述 4 种，唯密文攻击、已知明文攻击、选择明文攻击和选择密文攻击。除唯密文攻击外，其他 3 种攻击的目的是获得密钥。而对以一次一密为原型的密钥流加密体制来说，获取密钥是没有意义的。因为密钥在每一次加密时都要发生变化。所以对于攻击者而言只有密文这一唯一的信息来源。攻击者关心的不是密钥而是明文。

编码方案的选择取决于两个因素，效率和安全性。由定理2可知，事件序列从其源信号中获取的信息量的多少与区间的数量成正比例关系。因此，单从分区数量角度考虑，如果无限地增加区间数量，则一次迭代就可以得足够的信息量生成流密钥。但是，混沌过程退化严重，导致对初始值的依赖敏感性降低，从而安全性降低。

3.1 {0,1}序列伪随机性判断

Golomb建议了广为接受的判断二进制伪随机序列随机性的3个公设：

(1) 在序列的一个周期内，0与1的个数接近相等，更精确地说，相差至多为1；

(2)在序列的一个周期内，长为1的游程占流程总数的$1/2$，长为2的游程占总数的$1/2^2$，长为i的游程占游程总数的$1/2^i$，且在等长的游程中0的游程个数和1的游程个数相等。连续n个0或1称作长度为n的游程；

(3)自相关函数(Auto-Correlation)如式(8)所示。K为正数。

$$C(\tau)=\sum_{i=0}^{N-1}(-1)^{a_i+a_{i+\tau}}=\begin{cases}N, & \text{if } \tau\equiv 0(\bmod N)\\ K, & \text{if } \tau\not\equiv 0(\bmod N)\end{cases} \tag{8}$$

文中所有对二进制序列伪随机性的验证都是利用Golomb的3个公设进行的。

3.2 编码位数实验分析

本文采用一维线性分段混沌映射，如式(13)所示，进行编码离散化。本节实验的目的是从伪随机性和生成效率上分别找出最合适的编码位数。针对公设(1)，进行统计实验，共进行1000次。图1显示了离散后二进制伪随机序列的数学期望随描述位数的关系。图中多条曲线是有代表性的实验结果。所有的伪随机序列的标准差最大为0.0135。因此，采用的编码位数对伪随机性的影响不大，也即无论采用多少位编码，所得到的二进制伪随机序列的随机性水平相当。值得说明的一点是，当采用18位编码时，多次实验的曲线相交于0.5004，离0-1均匀分布的数学期望0.5最近，并且在18附近标准差明显有收敛趋势。

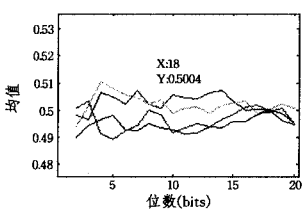


图1 均值与编码位数关系

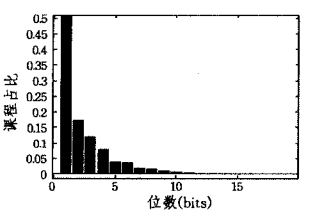


图2 游程比例与游程长度关系

针对公设(2)，对各长度的游程数量占总游程数量的平均比例进行了统计计算。结果如表1和图2所示。实验值是1000次实验中，特定长度的游程所占比例的平均效果。与标准值的误差在±0.01之内，符合伪随机性的要求。各长度的游程数量占比，都与图1类似，即编码位数对游程比例的影响也不明显。

针对公设(3)，各次实验产生的二进制序列都具有伪随机性。由上述分析可知，利用Golomb随机性公设可以对二进制序列的伪随机性进行判断，但是对确定编码位数帮助不大。

对编码位数与产生相等长度的二进制序列所需要的时间进行统计实验，得出平均效果图，如图3所示。

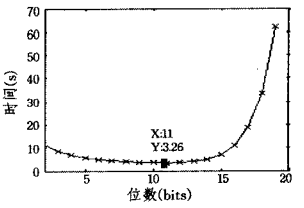


图3 离散过程用时与编码位数的关系

由图3分析可知，仅从效率角度选择编码位数时，编码位数在6位至15位之间离散过程所需要的时间差别不大，6位至15位均可。最低点在11位编码处。该区间之外，所需要的时间显著增加，并且增加的速率很大，生成效率以较大的斜率降低。

综上所述，编码法进行离散化产生的二进制序列在伪随机性上都较好并且有相当的表现，因此，从伪随机性无法确定采用多少位编码是合适的。编码位数对流密钥生成效率的影响是显著的，由效率确定编码位数是合理的。所以，最终结论是在产生的二进制序列具有伪随机性的前提下，由效率来决定编码位数。

表1 编码位数与特定长度的游程所占的比例的关系

类型	1	2	3	4	5	6	7	8
标准值	0.5000	0.2500	0.1250	0.0625	0.0313	0.0156	0.0078	0.0039
实验值	0.4979	0.2502	0.1340	0.0575	0.0336	0.0172	0.0024	0.0041

4 性能分析

4.1 理论分析

定理5 设有一明文P，其信息量为$I(P)$。设等间隔抽样的间隔数为n，则产生密钥流所用的迭代次数为$N=\lfloor I(P)/\log(n)\rfloor+1$ (9)

二域法的离散化过程最简单的办法是以混沌吸引子域的中点为分界线，将整个区域分成两个部分，也即有如式(10)所示的函数$T_n(\cdot)$，式中，l,h分别表示混沌吸引子的下界和上界的坐标。用1bit信息量描述每次迭代产生的点落在哪个区间。统计实验发现，由于混沌映射自身的原因，这种离散化过程容易造成离散化后的二进制伪随机序列并不服从均匀分布，因而发展了文献[7,8]中的离散化过程。函数$T_n(\cdot)$的表示如式(12)所示。式中，n为任意正整数；区间的分法与定理1相同。发展了的离散化过程等间隔地量化，实现了二进制伪随机序列的均匀化。文献虽然将区间分成了n个小区间，但是本质上并没有发生变化，依然只是用1bit信息量描述迭代的落点。

$$T_n(x)=\begin{cases}0, & \text{if } x\in[l,(h+l)/2)\\ 1, & \text{if } x\notin[l,(h+l)/2,h]\end{cases} \tag{10}$$

因此，理论上产生信息量为$I(P)$的流密钥需要的迭代次数为式(11)所示。

$$N=\lfloor I(P)\rfloor+1 \tag{11}$$

若采用多次混沌迭代来进行前馈，那么每次产生的有效信息量会更少，需要的迭代次数更多，生成效率也就更低。

因此，理论分析指出，在保证相当伪随机性的前提下，编码法的生成效率是二域法的编码位数倍。

4.2 实验及数据分析

4.2.1 实验用统计量

将X的可能取值的全体Ω分成k个两两不相交的子集

$A_1, A_2, \cdots, A_k$。以 $f_i(i=1,2,\cdots,k)$作为样本观察值 $x_1, x_2, \cdots, x_n$ 中落在 A_i 中的个数。

$$T_n(x)=\begin{cases}0, & \text{if } x\in\bigcup_{d=0}^{2^{n-1}-1}A_{2d}^n\\ 1, & \text{if } x\notin\bigcup_{d=0}^{2^{n-1}-1}A_{2d}^n\end{cases} \tag{12}$$

则有统计量

$$\sum_{i=1}^{k}h_i\left(\frac{f_i}{n}-p_i\right) \tag{13}$$

式中，$h_i(i=1,2,\cdots,k)$是给定的常系数。根据 Karl Pearsom 的证明，当取 $h_i=n/p_i$ 时，则式(13)近似地服从 $\chi^2(k-1)$分布。于是可得拒绝域为：

$$\chi^2\geqslant\chi_\alpha^2(k-1) \tag{14}$$

又当 k 比较大时，可由 R. A. Fisher 公式：

$$\chi_\alpha^2\approx\frac{1}{2}(z_a+\sqrt{2n-1})^2 \tag{15}$$

计算出拒绝域。其中，z_α 表示标准正态分布的分位点。

4.2.2 混沌映射函数的比较

实验采用混沌吸引子统计分布比较好的一维线性分段函数[10]，如式(16)所示。

$$x(t+1)=F_p(x_t)=\begin{cases}x(t)/p, & 0\leqslant x(t)<p\\ \dfrac{(x(t)-p)}{(0.5-p)}, & p\leqslant x(t)<0.5\\ F_p(1-x(t)), & x(t)\geqslant 0.5\end{cases} \tag{16}$$

该混沌映射的分布比前期常用的 logistic 混沌映射的分布在均匀性方面要好，如式(17)所示。

$$\begin{aligned}&x_{k+1}=\mu x_k(1-x_k),\\ &x_k\in(0,1),3.571\cdots\leqslant\mu\leqslant 4\end{aligned} \tag{17}$$

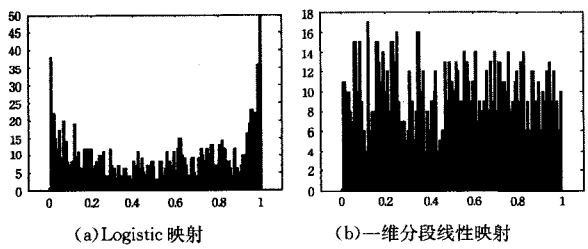
随机设定 1000 个初始值，每个初始值迭代 1000 次。按 4.2.1 节中介绍的 χ^2 拟合检验法分析数据，计算结果并取平均值如表 1 所列。取 $k=100$ 时，结果如表 2 所列。

表 2 分布统计量平均值

类型	Logistic	一维分段线性
拒绝域	≥135.001	
χ^2 值	531.001	95.800

Logistic 映射在显著性水平 0.01 下，其混沌分布不是均匀分布。而一维分段线性混沌映射可以认为是均匀分布。从 1000 次实验中各取有代表性的一个分布，如图 4 所示。

根据上面的分析，我们在实验中采用了式(16)所示的一维线性分段混沌映射函数。



(a)Logistic 映射　　(b)一维分段线性映射

图 4 两个映射的典型混沌分布直方图

4.2.3 效率比较

采用 8 位编码，也即将[0,1]区间分为 $k=256$ 个小区间时，分别用二域法和抽样编码法进行离散化。表 3 显示了重要指标的对比情况。采用 8 位编码时，二域法所需的时间约是抽样编码法所需要时间的 8 倍。在不同机器上实验时，绝对时间不同，但是所用时间的比例关系保持稳定。

表 3 二域法与抽样编码法各项指标的比较

方法	数学期望	平均用时(秒)
二域法	0.4948	24.6331
抽样编码法	0.4981	3.0458

综上所述，抽样编码法的效率约是二域法的 8 倍，与理论分析结果相符。编码法能有效提高流密钥的生成效率。

结束语 采用信息论的方法描述了混沌密钥流的生成过程。从伪随机性信息的传递角度来分析二进制密钥流的生成过程。提出了抽样编码离散化方法，用来转化伪随机性信息。理论推导和实验证明了方案在保证伪随机性的条件下，能明显地提高生成流密钥的效率。研究了编码位数对伪随机性和流密钥生成效率的影响，得出在保证伪随机性的前提下由生成效率决定编码位数的结论。在理论和实验两个方面比较了二域法和编码法，证明了编码法能有效提高生成效率。

已有文献中常采用的 Logistic 混沌映射函数的伪随机性没有文中所采用的一维线性分段混沌映射函数的伪随机性好。但是，已有文献证明后者存在线性缺陷，即已知两个或更多的点落在同一分段上，容易造成已知部分明文攻击。采用本文提出的抽样编码方法将混沌吸引子分成更多的小区间，在一定程序上降低了线性缺陷发生的概率。本文的重点是验证采用抽样编码方法进行二进制序列生成能否保证伪随机性。所以，准确分析线性缺陷发生的概率，采用其他伪随机性更好、又没有线性缺陷的混沌映射函数是进一步要完成的工作。

参 考 文 献

[1] 包浩明，朱义胜. 基于多层密钥的混沌映射保密通信系统[J]. 电子学报，2009，37(6)：1222-1125

[2] 芦斌，罗向阳，刘粉林. 一种基于混沌的软件水印算法框架及实现[J]. 软件学报，2007，18(2)：351-360

[3] 张瀚，王秀峰，李朝晖，等. 一种基于混沌系统及 Henon 映射的快速图像加密算法[J]. 计算机研究与发展，2005，42(12)：2137-2142

[4] 王培荣，徐喆，付冲，等. 复合混沌数字图像加密算法[J]. 通信学报，2006，27(11)：285-289

[5] Habutsu T，Nishio Y，Sasase I，et al. A Secret Key Cryptosystem by Iterating a Chaotic Map[C]// Advanced in cryptology-EOROCRYPT'91. Workshop on the Theory and Application of Cryptographic Techniques. Brighton，UK，Berlin：Springer Heidelberg，1991，547：127-136

[6] 王相生. 序列密码设计与实现的研究[D]. 上海：中国科学院上海冶金研究所，2001

[7] 周红，罗杰，凌燮亭. 混沌非线性反馈密码序列的理论设计和有限精度实现[J]. 电子学报，1997，25(10)：57-60

[8] 桑涛，王汝笠，严义埙. 一类新型混沌反馈密码序列的理论设计[J]. 电子学报，1999，27(7)：47-50

[9] 李红达，冯登国. 基于复合离散混沌动力系统的序列密码算法[J]. 软件学报，2003，14(5)：991-998

[10] 周红，罗杰，凌燮亭. 混沌前馈型流密码的设计[J]. 电子学报，1998，26(1)：98-101

[11] Jakimoski G，Kocare L. Chaos and Cryptography-part II：Block Encryption Based on Chaotic Maps[J]. Circuits and Systems I：Fundamental Theory and Applications，IEEE Transactions on，2001，48(2)：163-169