一种兼具对称加密和无损压缩双重功能的编码算法

王杰林[1,2]，廖亦凡[1]

（1 湖南涉外经济学院信息与机电工程学院，湖南 长沙 410205；
2 湖南遥昇通信技术有限公司，湖南 长沙 410600）

摘　要：无损压缩算法无法实现高强度对称加密，对称加密算法无法实现良好的压缩效果，使得系统中无损压缩算法和对称加密算法需级联应用，需两种算法的硬件资源和运算能耗。算法内联应用可减少硬件资源和运算能耗。本文定义了加权概率质量函数和加权分布函数，提出了一种基于加权概率模型无损编码方法并分析了该算法的信息熵。构造了加权系数的非线性轮函数、分段交换编码，结合密码的非线性异或运算给出了兼具对称加密和无损压缩双重功能的编码算法，从数学模型上实现算法内联。实验结果表明：本文算法可实现良好的无损压缩效果，又兼具高强度的数据加密。该算法简单易硬件实现，可广泛应用于通信、存储和安全系统。

关键词：对称加密；无损压缩；加权概率模型

中图分类号：TP309.7

文献标志码：A

A coding algorithm with dual functions of symmetric encryption and lossless compression

WANG Jielin[1, 2], LIAO Yifan[1]

1 School of information and mechanical and electrical engineering, Hunan International Economics University, Changsha, 410205, China

2 Hunan YESINE Communication Technology Co., Ltd., Changsha, 410600, China

Abstract: The lossless compression algorithm cannot achieve high-strength symmetric encryption, and the symmetric encryption algorithm cannot achieve a good compression effect. Therefore, the lossless compression algorithm and the symmetric encryption algorithm in the system need to be applied independently in cascade. At present, some scholars study the inline application of algorithms, but they cannot solve the logic conflicts of algorithms well. This paper defines the weighted probability mass function and weighted distribution function, proposes a lossless coding method based on the weighted probability model, and analyzes the information entropy of the algorithm. The non-linear round function of the weighting coefficient and the segmented cross-coding are constructed, the algorithm implementation steps based on the non-linear XOR operation of the password information are given and the experimental test is carried out. The results show that the algorithm in this paper can achieve good lossless compression effects, and has high-strength symmetric encryption. It is a coding algorithm with dual functions of symmetric encryption and lossless compression. The algorithm is simple and easy to implement in hardware, and can be widely used in communication, storage and security systems.

Keywords: symmetric encryption; lossless compression; weighted probability model

1　引言

无损压缩算法（熵编码[1][2]）已经被广泛应用于通信、存储等领域，常见的无损压缩算法有行程编码，字典编码[3]，哈夫曼编码[4]以及算术编码（区间编码）[5][6][7]等。对称加密算法[8]作为信息安全的核心工具，也被广泛应用于通信、交易、支付以及数据脱敏等领域，常见的对称加密算法有 DES[9]，AES[10][11]，Blowfish[12]和 RC[13]等。在系统应用中，无损压缩算法无法实现高强度对称加密，对称加密算法无法实现良好的压缩效果，对称加密和无损压缩均为信源编码，这类算法一般采用独立级联应用[14][15][16][17]，因数据需通过先后两种算法的编译码，所以需两种算法的硬件资源和运算能耗。大数据环境下，对称加密和无损压缩采用算法内联可减少硬件资源和运算能耗。基于算术编码加密方法[18][19][20]是一种算法内联方法，然而该方法编码某一符号时通常映射成另一符号，因映射方法确定，所以必然改变了原有数据的信息熵，使得压缩效果无法确定。

本文基于概率分布函数[21]定义了加权概率分布函数和加权概率模型，分析加权系数与秘钥关系，构造加权系数的非线性轮函数和分段交换编码算法。该算法是一种新的数据无损编码方法，通过加权系数轮函数和分段交换编码实现数据的无损压缩和加密，从数学模型上实现算法的内联。因分段交换编码不会改变数据中各符号的概率，加权系数为编码时的计算变量同样不改变数据中各符号的概率，使得数据的信息熵不会改变。相较于目前主流的熵编码和对称加密算法，本文算法无损压缩可达信息熵，兼具高强度的

通信作者：廖亦凡, E-mail: 36476379@qq.com
基金项目：湖南省教育厅重点科研项目（湘教通[2019]353 号：19A283）。

对称加密，能有效降低硬件资源和运算能耗。

2 加权概率模型编码算法和证明

2.1 加权概率模型编码

令信源序列$X=(X_1,X_2,\dots,X_i,\dots,X_n)$是有限个值或可数个可能值的离散序列，$X_i\in A=\{0,1,2,\dots,k\}$。于是对于$A$中一切数值有概率空间：

$$\begin{bmatrix}X_i\\P\end{bmatrix}=\begin{bmatrix}0 & 1 & \dots & k\\ p(0) & p(1) & \dots & p(k)\end{bmatrix}$$

由于随机过程必须转移到某个符号，所以在任意时刻有：

$$\sum_{X_i=0}^{k}p(X_i)=1,\qquad 0\le p(X_i)\le 1$$

于是，任意符号X_i的分布函数为：

$$F(X_i)=\sum_{s\le X_i}p(s) \tag{1}$$

$p(0)\le F(x)\le 1,\ s\in A$。

定义 1 设离散随机变量X，$X\in A=\{0,1,\dots,k\}$，$\mathrm{P}\{X=a\}=p(a)(a\in A)$，加权概率质量函数为$\varphi(a)=r\mathrm{P}\{X=a\}=rp(a)$，$p(a)$为的概率质量函数，$0\le p(a)\le 1$，$r$为权系数，且

$$F(a)=\sum_{i\le a}p(i) \tag{2}$$

若$F(a,r)$满足$F(a,r)=rF(a)$，则称$F(a,r)$为加权累积分布函数，简称加权分布函数。显然，所有符号的加权概率之和为$\sum_{a=0}^{k}\varphi(a)=r$。

令离散信源序列$X=(X_1,X_2,\dots,X_n)$，$X_i\in A$，且令$F(X_i-1)=F(X_i)-p(X_i)$，序列X的加权分布函数记为$F(X,r)$。当$n=1$时：

$$F(X,r)=rF(X_1-1)+rp(X_1)$$

当$n=2$时：

$$F(X,r)=rF(X_1-1)+r^2F(X_2-1)p(X_1)+r^2p(X_1)p(X_2)$$

当$n=3$时：

$$F(X,r)=rF(X_1-1)+r^2F(X_2-1)p(X_1)+r^3F(X_3-1)p(X_1)p(X_2)+r^3p(X_1)p(X_2)p(X_3)$$

令$\prod_{j=1}^{0}p(X_j)=1$，类推得：

$$F(X,r)=\sum_{i=1}^{n}r^iF(X_i-1)\prod_{j=1}^{i-1}p(X_j)+r^n\prod_{i=1}^{n}p(X_i) \tag{3}$$

将满足(3)的加权分布函数的集合定义为加权概率模型，简称加权模型，记为$\{F(X,r)\}$。若$X_i\in A=\{0,1\}$，则称$\{F(X,r)\}$为二元加权模型。令：

$$H_n=F(X,r) \tag{4}$$

$$R_n=rp(X_1)rp(X_2)\dots rp(X_n)=r^n\prod_{i=1}^{n}p(X_i) \tag{5}$$

$$L_n=H_n-R_n \tag{6}$$

其中$X_i\in A,n=1,2,\dots$。当$r=1$时：

$$F(X,1)=\sum_{i=1}^{n}F(X_i-1)\prod_{j=1}^{i-1}p(X_j)+\prod_{i=1}^{n}p(X_i) \tag{7}$$

由(4)(5)(6)可得$H_n=F(X,1)$，即区间编码（算术编码）是基于$r=1$时加权分布函数的无损编码方法。

因X_i必须取A中的值，所以$p(X_i)>0$。显然(4)(5)(6)为区间列，$[L_i,H_i)$是信源序列X在时刻$i(i=0,1,2,\dots,n)$变量X_i对应的区间上下标，$R_i=H_i-L_i$是区间的长度。根据(4)(5)(6)，设$i=0$时$R_0=H_0=1$，$L_0=0$，于是$i=1,2,\dots,n$时加权概率模型编码运算式为：

$$\begin{aligned} R_i &= R_{i-1}\varphi(X_i) \\ L_i &= L_{i-1} + R_{i-1}F(X_i - 1, r) \\ H_i &= L_i + R_i \end{aligned} \tag{8}$$

通过(8)对信源序列X进行加权概率模型编码运算，L_n为实数，是加权概率模型编码结果。L_n通过进制转换得到二进制序列。以二进制序列为例，令$0 < r \le 1$且序列X从$i+1$位置开始的 3 个符号为0,1,0。根据(8)加权模型的编码运算过程如图 1。

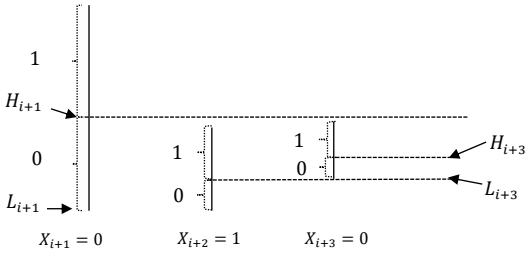

图 1 加权模型编码 010 的过程示意图

根据图 1，若$H_{i+3} > H_{i+1}$，因区间$[H_{i+1}, H_{i+3}) \in [H_{i+1}, H_{i+1} + R_{i+1})$，且$[H_{i+1}, H_i + R_i)$与符号 1 对应，所以第$i+1$个符号 0 可能被错误译码为符号 1。若$H_{i+3} \le H_{i+1}$，则$[L_{i+3}, H_{i+3}) \in [L_{i+1}, H_{i+1})$。如图 1 中$[L_{i+1}, H_{i+1})$与符号 0 唯一对应，所以$i+1$位置上的符号 0 被$L_{i+3}$正确译码，且$i+2$和$i+3$位置上的符号 1 和符号 0 也能正确译码。当$0 < r \le 1$时，任意时刻都有$[L_{i+1}, H_{i+1}) \in [L_i, H_i)$，可无损译码。

2.2 无损译码证明

定理 2 加权模型满足：

⑴ $L_n < H_n \wedge L_n < H_{n-1} \wedge \ldots \wedge L_n < H_1$，通过$L_n$可完整还原序列$Q$；

⑵ $\lim\limits_{n\to\infty}(H_n - L_n) = 0$，即收敛性；

⑶ $\lim\limits_{n\to\infty} H_n = L_n$，即唯一性。

证明 ⑴根据(8)，L_n为单调不减函数，当且仅当$L_n \in [L_n, H_n) \wedge L_n \in [L_{n-1}, H_{n-1}) \wedge \ldots \wedge L_n \in [L_1, H_1)$时，因$[L_i, H_i)(i = 1,2,\ldots,n)$与变量$X_i$为唯一映射关系，所以当$L_n \in [L_i, H_i)(i = 1,2,\ldots,n)$时得出唯一的符号$X_i$，从而完整得出信源序列$X$，于是$L_n < H_n \wedge L_n < H_{n-1} \wedge \ldots \wedge L_n < H_1$。

⑵根据(5)，$0 < r \le 1$且$0 \le p(X_i) \le 1$，当信源序列$X = \{X_i = a\}$且$r = 1$时$p(a) = 1$，$R_n = 1$，$L_n \to F(a, r) = H_n$，于是$H_n - L_n \to 0$。当$0 < p(X_i) < 1$且$0 < r \le 1$时$0 < rp(X_i) < 1$，$R_n \to 0$，因$H_n - L_n = R_n$，所以$H_n - L_n \to 0$。所以$n \to \infty$时$\lim\limits_{l\to\infty}(H_l - L_l) = \lim\limits_{l\to\infty} R_l = 0$，加权概率模型是收敛的。

⑶$\{L_n\}$是严格单调不减且有上界的数列，由单调有界定理，设$\lim\limits_{n\to\infty} L_n = \xi$，且$\xi \ge L_n$。因为$\lim\limits_{n\to\infty}(H_n - L_n) = 0$，所以$\lim\limits_{n\to\infty} L_n = \lim\limits_{n\to\infty} H_n = \xi$，所以$\xi = L_n$，$\lim\limits_{n\to\infty} H_n = \xi = L_n$，且$L_n$是唯一的。

3 加权概率模型信息熵和安全加密方案

3.1 加权概率模型信息熵

设离散无记忆信源序列$X = (X_1, X_2, \ldots, X_n)(X_i \in A, A = \{0,1,2,\ldots,k\})$，当$r = 1$时，$\varphi(X_i) = p(X_i)$。由香农信息熵[1][2]定义，$X$的熵为：

$$H(X) = -\sum_{X_i=0}^{k} p(X_i)\log_{k+1} p(X_i) \tag{9}$$

当$r \ne 1$时，定义具有概率$\varphi(X_i)$的随机变量X_i的自信息量为：

$$I(X_i) = -\log_{k+1} p(X_i) \tag{10}$$

设集合$\{X_i = a\}(i = 1,2,\ldots,n, a \in A)$中有$c_a$个$a$。当$r$的值确定，信源序列$X$的总信息量为：

$$-\sum_{a=0}^{k} c_a \log_{k+1} p(a)$$

于是平均每个符号的信息量为：

$$-\frac{1}{n}\sum_{a=0}^{k} c_a \log_{k+1} p(a) = -\sum_{a=0}^{k} p(a)\log_{k+1} p(a)$$

定义 3 令$H(X,r)$为：

$$\begin{aligned} H(X,r) &= -\sum_{a=0}^{k} p(a)\log_{k+1}\varphi(a) \\ &= -\log r - \sum_{a=0}^{k} p(a)\log_{k+1} p(a) \\ &= -\log r + H(X) \end{aligned} \tag{11}$$

根据定义 3，在r的值确定时，通过加权概率模型编码后的二进制长度为$nH(X,r)(bit)$。

3.2 非线性加权系数的轮函数

根据(8)可得序列X的权系数r。若r不随i（$i = 1,2,\dots,n$）变化，则称r为静态加权系数；若r随i变化，则称r为动态加权系数，记为$r(i)$。根据定理 2 可得当$0 < r \le 1$时加权概率模型可无损编译码，又根据定义 3 可得当$r \to 1$时，$-\log r \to 0$，于是$H(X,r) \to H(X)$。显然，基于动态加权系数$r(i)$，当$0 < r(i) \le 1$且$r(i) \to 1$时加权概率模型编码方法可达信息熵。由于$r(i)$与当前符号的序号i有关，且与每个符号的编码运算相关，符合秘钥的定义。因动态加权系数$r(i)$适合构造非线性轮函数，于是定义$r(i)$为加权概率模型对称加密秘钥。基于动态权系数$r(i)$的加权分布函数为：

$$F(X,r) = \sum_{i=1}^{n} F(X_i - 1)\prod_{j=1}^{i} r(j)\prod_{j=1}^{i-1} p(X_j) + \prod_{i=1}^{n} r(i)\prod_{i=1}^{n} p(X_i) \tag{12}$$

因$0 < r(i) < 1$且$0 \le p(X_j) \le 1$，不难得出$0 \le F(X,r) < 1$，编码后得到唯一的实数L_n，L_n可转换为m个字节序列Y。于是：

$$L_n = \sum_{i=1}^{n} F(X_i - 1)\prod_{j=1}^{i} r(j)\prod_{j=1}^{i-1} p(X_j) \tag{13}$$

常见 DES，AES，Blowfish 对称加密算法采用非线性字节替换和轮函数（混沌映射）来消除线性相关性，通常称为 S 盒。设输入的密码（Passwords）共L（$L \ge 6$）个字节，n为序列X的字节数，$B(j)$为密码第j（$j = 1,2,\dots,L$）个字节值，即$B_j = 0,1,\dots,255$。当$n < L$时$B(j)$经下式运算后使得密码中的每个字节均可有效作用于加权系数，于是有$L = n$。

$$B(j) = B(j) \oplus B(L - j) \tag{14}$$

随机生成T^2（$T = 16$）个 0 到 255 数值存于$T * T$的二维表中，如表 1。S 盒中第t（$t = 1,2,\dots,T^2$）个字节值$S(t)$计算方法为：

$$S(t) = B(t \bmod L) \oplus B\big((t+1) \bmod L\big) \oplus f(t) \tag{15}$$

$f(t)$表 1 中第t个字节值。显然仅表 1 已知，B未知，则$S(t)$不可知，仅正确的密码方能得出正确坐标和 S 盒。经(15)计算后 S 盒中的值可能存在相同的值，无法实现类 AES 的字节代换。令 S 盒中坐标为(x,y)的字节值为 $g(x,y)$，x和y的计算式为：

$$x = (X_i \oplus B(i \bmod L)) \bmod T, \qquad y = (X_i \oplus B\big(L - (i \bmod L)\big)) \bmod T \tag{16}$$

其中X_i为序列X第i个字节值，定义$r(i)$的非线性轮函数为：

$$r(i) = 1 - \frac{B\Big(\big(B(i \bmod L) \oplus B\big((i+1) \bmod L\big)\big) \bmod L\Big)256 + \ g(x,y)}{10^s} \tag{17}$$

其中$\big(B(i \bmod L) \oplus B\big((i+1) \bmod L\big)\big) \bmod L$为密码字节序列的下标。 $s \ge 6$，s的实际值根据计算机的运算精度而定，本文实验中$s = 7$。s越大，则$r(i)$越趋近于 1，于是加权概率模型编码方法可接近信息熵，因为$r(i) \to 1$，$-\log r \to 0$，则$H(X,r) \to H(X)$。因$r(i)$为第i个符号X_i编码时R_i和迭代运算的权系数，即：

$$\begin{aligned} R_i &= R_{i-1}\varphi(X_i) = R_{i-1} r(i) p(X_i) \\ L_i &= L_{i-1} + R_{i-1} r(i) F(X_i - 1) \end{aligned} \tag{18}$$

由(18)可得权系数轮函数是通过R_i和L_i迭代，迭代次数为n。(16)和(17)使得$r(i)$、$r(i+1)$和$r(i-1)$不存在线性关系，因为$B\Big(\big(B(i \bmod L) \oplus B\big((i+1) \bmod L\big)\big) \bmod L\Big) \in \{0,1,\dots,255\}$且$g(x,y) \in \{0,1,\dots,255\}$，所以$r(i) \in R = \{1.0 - \frac{65535}{10^s}, 1.0 - \frac{65534}{10^s}, 1.0 - \frac{65533}{10^s}, \dots, 1.0\}$，当$s$确定，集合$R$有$65536$个实数值，可得$r(i)$取集合$R$中任意实数值的概率为$\frac{1}{65536}$。

当X和Y已知，因Y已知则L_n已知，所以L_1，L_2，…， L_{n-1}均存在一个趋势，即均为小于L_n且接近L_n的值。$r(1)$有65536个可能值。 $p(X_1)$和$F(X_1 - 1)$已知且$R_0 = 1$，$L_0 = 0$，由(18)可得L_1仅与$r(1)$有关，所以L_1有65536个可能值，找出小于L_n且接近L_n的值。又因X_2已知则$p(X_2)$和$F(X_2 - 1)$已知，于是L_1和L_2同时满足小于L_n且接近L_n的可能值数量进一步减少，很容易逼近或计算出密码。

若X_1未知，因$p(0)$，$p(1)$，…，$p(255)$已知，所以$F(X_1-1)$存在256个可能值（当$X_1=0$时$F(-1)=0$）。因$r(1)$存在65536个可能值，$r(1)$与$F(X_1-1)$为相互独立变量，所以L_1存在256^3个可能值。若X_1和X_2未知，$F(X_2-1)$存在256个可能值，且$r(2)$存在65536个可能值。因$L_2=L_1+R_0r(1)p(X_1)r(2)F(X_2-1)$，所以仅需考虑$r(1)$、$p(X_1)$、$r(2)$和$F(X_2-1)$四个相互独立变量，可得$L_2$存在$256^6$个可能值。但是当$Y$已知时，通过$L_1$和$L_2$是否满足“小于$L_n$且接近$L_n$”去除无效值。若$Y$未知，则$L_1$和$L_2$失去关键条件“小于$L_n$且接近$L_n$”。类推可得$L_n$存在$256^{3n}$个可能值。

上述分析攻击可能性，加权模型编码具备无损压缩作用，则序列X各符号的概率不可变，经下面两个方法实现高强度加密作用。

1、编码明文前，预先编码u个随机字节，将牺牲一定的压缩比。因随机字节未知，所以X不可知。编码序列长度增加，L_{n+u}存在$256^{3(n+u)}$个可能值。

2、将序列Y中的字节与 S 盒中的字节异或运算得出密文，因序列Y为编码后的字节序列，所以不影响压缩比。译码时 S 盒由密码和表 1 动态生成，因密码未知，所以 S 盒和(x,y)不可知，所以Y不可知。设Y_i为编码后序列Y的第i（$i=1,2,\dots,m$）个字节，Y_i与$S(i \bmod T^2)$运算式为：

$$Y_i=Y_i\oplus S(i \bmod T^2) \tag{19}$$

表 1 随机生成T^2个 0 到 255 不重复整数的二维表

x \ y	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
0	0x63	0x7c	0x77	0x7b	0xf2	0x6b	0x6f	0xc5	0x30	0x01	0x67	0x2b	0xfe	0xd7	0xab	0x76
1	0xca	0x82	0xc9	0x7d	0xfa	0x59	0x47	0xf0	0xad	0xd4	0xa2	0xaf	0x9c	0xa4	0x72	0xc0
2	0xb7	0xfd	0x93	0x26	0x36	0x3f	0xf7	0xcc	0x34	0xa5	0xe5	0xf1	0x71	0xd8	0x31	0x15
3	0x04	0xc7	0x23	0xc3	0x18	0x96	0x05	0x9a	0x07	0x12	0x80	0xe2	0xeb	0x27	0xb2	0x75
4	0x09	0x83	0x2c	0x1a	0x1b	0x6e	0x5a	0xa0	0x52	0x3b	0xd6	0xb3	0x29	0xe3	0x2f	0x84
5	0x53	0xd1	0x00	0xed	0x20	0xfc	0xb1	0x5b	0x6a	0xcb	0xbe	0x39	0x4a	0x4c	0x58	0xcf
6	0xd0	0xef	0xaa	0xfb	0x43	0x4d	0x33	0x85	0x45	0xf9	0x02	0x7f	0x50	0x3c	0x9f	0xa8
7	0x51	0xa3	0x40	0x8f	0x92	0x9d	0x38	0xf5	0xbc	0xb6	0xda	0x21	0x10	0xff	0xf3	0xd2
8	0xcd	0x0c	0x13	0xec	0x5f	0x97	0x44	0x17	0xc4	0xa7	0x7e	0x3d	0x64	0x5d	0x19	0x73
9	0x60	0x81	0x4f	0xdc	0x22	0x2a	0x90	0x88	0x46	0xee	0xb8	0x14	0xde	0x5e	0x0b	0xdb
10	0xe0	0x32	0x3a	0x0a	0x49	0x06	0x24	0x5c	0xc2	0xd3	0xac	0x62	0x91	0x95	0xe4	0x79
11	0xe7	0xc8	0x37	0x6d	0x8d	0xd5	0x4e	0xa9	0x6c	0x56	0xf4	0xea	0x65	0x7a	0xae	0x08
12	0xba	0x78	0x25	0x2e	0x1c	0xa6	0xb4	0xc6	0xe8	0xdd	0x74	0x1f	0x4b	0xbd	0x8b	0x8a
13	0x70	0x3e	0xb5	0x66	0x48	0x03	0xf6	0x0e	0x61	0x35	0x57	0xb9	0x86	0xc1	0x1d	0x9e
14	0xe1	0xf8	0x98	0x11	0x69	0xd9	0x8e	0x94	0x9b	0x1e	0x87	0xe9	0xce	0x55	0x28	0xdf
15	0x8c	0xa1	0x89	0x0d	0xbf	0xe6	0x42	0x68	0x41	0x99	0x2d	0x0f	0xb0	0x54	0xbb	0x16

4 二进制加权概率模型多功能编码算法实现

4.1 编码算法流程

设输入的密码共L（$L\geq 6$）个字节序列B，$B(i)$为序列B第i个字节，采用加权概率模型对长度为n的字节序列X编码步骤如下。

（1）初始化参数，$L_0\leftarrow 0$，$H_0\leftarrow R_0\leftarrow 1$，$i\leftarrow j\leftarrow 0$；$X_0=0$；

（2）定义 S 盒数组$S[T^2]$；

（3）统计序列X符号 0 的个数c，$p\leftarrow\frac{c}{n}$；

（4）随机生成长度为u（$u\geq 16$）的字节序列Q；

（5）$n\leftarrow n+u$；

（6）当$n\geq L$时，转（9）；

（7）当$i<n$时$B(i)\leftarrow B(i)\oplus B(L-i)$，$i\leftarrow i+1$，重复（7）；

（8）$i\leftarrow 0$，$L\leftarrow n$；

（9）根据j查表 1，得出$f(j)$；

（10）当$j<T^2$时，$S[j]\leftarrow B(j \bmod L)\oplus B\big((j+1)\bmod L\big)\oplus f(j)$且$j\leftarrow j+1$，重复（9）到（10）；

（11）串联Q（前）和X（后）得到二进制序列Z；

（12）获取序列Z第i个字节值X_i；

（13）获取密码第$(i \bmod L)$个字节值$B(i \bmod L)$和第$L-(i \bmod L)$个字节$B\big(L-(i \bmod L)\big)$；

（14）计算x和y，$x\leftarrow (X_{i-1}\oplus B(i \bmod L)) \bmod T$，$y\leftarrow (X_{i-1}\oplus B\big(L-(i \bmod L)\big)) \bmod T$；

（15）根据$xT+y$查 S 盒获取$g(x,y)$；

（16）根据式(17)计算$r(i)$；

（17）计算加权概率$\varphi(0)$和$\varphi(1)$，$\varphi(0)\leftarrow r(i)p$，$\varphi(1)\leftarrow r(i)(1-p)$；

（18）若$X_i=0$，$R_i \leftarrow R_{i-1}\varphi(0)$，否则$L_i \leftarrow L_{i-1}+R_{i-1}\varphi(0)$，$R_i \leftarrow R_{i-1}\varphi(1)$；

（19）$i \leftarrow i+1$，若$i<n$，重复（12）到（19）；

（20）将L_n转换成字节序列Y，将Y的长度记为m，$i \leftarrow 0$；

（21）获取序列Y第i个字节值Y_i；

（22）当$i<m$时，$Y_i \leftarrow Y_i \oplus S(i \bmod T^2)$，$i \leftarrow i+1$，重复（21）（22）；

（23）输出Y，m，n，c，u，结束编码。

4.2 译码算法流程

4.1 节基于式(8)编码字节序列X第i个符号X_i。当$X_i=0$时$F(0-1,r)=rF(-1)$，因$-1 \notin \{0,1,\dots,k\}$，所以$F(-1)=0$。

$$L_i=L_{i-1},R_i=R_{i-1}\varphi(0)$$

当$X_i=a$且$a=1,2,\dots,k$时：

$$L_i=L_{i-1}+R_{i-1}F(a-1,r),R_i=R_{i-1}\varphi(a)$$

根据定理 2，无损译码条件为$L_n \in [L_n,H_n) \wedge L_n \in [L_{n-1},H_{n-1}) \wedge \dots \wedge L_n \in [L_1,H_1)$，区间$[L_i,H_i)$与符号$X_i$一一映射。若$L_i \geq L_{i-1}+R_{i-1}F(a-1,r)$时$X_i=a$，于是译码$X_i$时，令

$$H(a)=L_{i-1}+R_{i-1}F(a-1)$$

则$L_n \geq H(a)$时$X_i=a$。

设输入的密码字节序列B'共M个字节，译码端获得Y，m，n，c和u，采用加权概率模型译码二进制信源序列X步骤如下：

（1）初始化参数，$L_0 \leftarrow 0$，$R_0 \leftarrow 1$，$H \leftarrow 1$，$i \leftarrow j \leftarrow 0$；$X_0=0$；

（2）计算符号 0 的概率，$p \leftarrow \frac{c}{n}$；

（3）定义 S 盒数组$S[T^2]$；

（4）当$n \geq M$时，转（7）；

（5）当$i<n$时$B'(i) \leftarrow B'(i) \oplus B'(M-i)$，$i \leftarrow i+1$，重复（5）；

（6）$i \leftarrow 0$，$M \leftarrow n$；

（7）根据j查表 1，得出$f(j)$；

（8）当$j<T^2$时，$S[j] \leftarrow B'(j \bmod M) \oplus B'\big((j+1) \bmod M\big) \oplus f(j)$且$j \leftarrow j+1$，重复（7）到（8）；

（9）获取序列Y第i个字节值Y_i；

（10）当$i<m$时，$Y_i \leftarrow Y_i \oplus S(i \bmod T^2)$，$i \leftarrow i+1$，重复（9）（10）；

（11）将字节序列Y转换成L_n，$i \leftarrow 0$；

（12）获取密码第$(i \bmod L)$个字节值$B'(i \bmod L)$和第$L-(i \bmod L)$个字节$B'\big(L-(i \bmod L)\big)$；

（13）计算x和y，$x \leftarrow (X_{i-1} \oplus B'(i \bmod L)) \bmod T$，$y \leftarrow (X_{i-1} \oplus B'\big(L-(i \bmod L)\big)) \bmod T$；

（14）根据$xT+y$查 S 盒获取$g(x,y)$；

（15）根据式(17)计算$r(i)$；

（16）计算全部符号的加权概率$\varphi(a) \leftarrow r(i)p(a)$，其中$a=0,1,2,\dots,k$；

（17）计算全部符号的加权分布函数$H(a)$，$H(a) \leftarrow L_i+R_iF(a-1)$，其中$a=0,1,2,\dots,k$；

（18）若$L_n \geq H(a)$，则第i个符号为$X_i=a$，$L_i \leftarrow L_i+R_iF(a-1)$，$R_i \leftarrow R_{i-1}\varphi(a)$；

（19）$i \leftarrow i+1$，若$i<n$，重复（12）到（19）；

（20）去除前u个字节，结束译码。

4.3 实验和分析

（1）本文方法与 DES、AES 的比较

DES 和 AES 采用 128 位秘钥的 ECB（ECB, Electronic Codebook Book）模式。密码设为 12345678，仿真实验一使用 Lena.bmp（Lena.bmp 是检验和测试图像压缩的常用素材，786486 字节 24 位真彩色），实验二随机生成符号等概率长度为 5242880 字节的数据（重复 100 次，得出编码后平均字节数），实验三为随机生成符号 0 的概率略等于$p=0.1$长度为 5242880 比特的二进制数据。得出实验数据如下。

表 2 Lena.bmp 仿真结果

算法	编码后的字节长度（字节）
本文算法	728624

DES	786488
AES	786488

表 3 随机字节数据仿真结果

算法	编码后的字节长度（字节）
本文算法	5242893
DES	5242888
AES	5242888

表 4 随机二进制数据仿真结果

算法	编码后的字节长度（字节）
本文算法	9092
DES	5242896
AES	5242896

实验中 DES 和 AES 编码后文件均大于原始文件。本文算法编码随机数据可接近信息熵，编码 Lena.bmp 和随机二进制数据时具有压缩效果。

（2）本文方法与算术编码的比较

算术编码已经广泛应用于 H264/H265 视频压缩标准[22][23]、图像[24]及文档压缩领域。实验一使用 Lena.bmp，实验二随机生成符号等概率长度为 5242880 字节数据（重复 100 次，得出编码后平均字节数），实验三为随机生成符号 0 的概率略等于$p=0.1$长度为 5242880 的二进制数据。密码以字节方式传递给算法程序，密码同上。

表 5 Lena.bmp 仿真结果

算法	编码后的字节长度（字节）
本文算法	727956
算术编码	727936

表 6 随机字节数据仿真结果

算法	编码后的字节长度（字节）
本文算法	5242912
算术编码	5242886

表 7 随机二进制数据仿真结果

算法	编码后的字节长度（字节）
本文算法	9092
算术编码	9038

实验本文算法压缩比略低于算术编码，原因分别为$r(i)<1$和随机二进制序列Y。当$r(i)=1$时，加权概率模型编码方法为算术编码，压缩比相同。因$s=5$则$r(i)\to 1$，所以本文算法与算术编码压缩比接近。译码时，输入的密码正确则可无损译码二进制序列X，否则译码出错误的二进制序列。经统计，本文算法编码后的二进制序列中符号趋于均等。

4.4 安全分析

4.2 节译码时需输入密码字节序列B'，且B'共M个。显然，当$B'=B$且$M=L$时可无损译码。

当$B'\neq B$且$M=L$，或者$B'\neq B$且$M\neq L$，译码步骤（7）查表 1 得出$f(j)$的值错误，于是步骤（8）得出的 S 盒值错误，步骤（9）和（10）得出的序列Y错误，则步骤（11）中L_n错误。因步骤（14）中$g(x,y)$错误，所以步骤（15）中$r(i)$错误，于是步骤（16）（17）（18）中$\varphi(a)$和$H(a)$错误，又因前u个随机字节未知，无法确定符号X_i是否译码正确。

编译码过程基于密码和动态生成的 S 盒，且基于随机的 S 盒和随机坐标(x,y)计算出$r(i)$进行加权概率模型编码，独立于待加密的数据。所以待加密的数据线性相关性和代数关系均无法作为本文方法破解的参考因素。尽管表 1 已知，但 S 盒未知，且随机坐标(x,y)未知，无法基于表 1 得出 S 盒或坐标(x,y)。基于 3.2 节的分析，未知密码的情形下，准确译码出前u个字节的概率为$\frac{1}{256^{3u}}$。当X已知，首先需进行256^{3u}次遍历，才能基于X_1开始分析密码的字节信息。但因序列Y未知，所以L_n不能作为译码时L_i的参照对象，无法确定密码的字节信息是否推测正确，所以线性译码时无法得出密码信息。可以得出本文的方法是安全的。

5 结束语

本文基于加权概率模型，构造了兼具对称加密和无损压缩双重功能的无损编码算法，从数学模型上实现算法内联。该算法无损

压缩可达信息熵，具备高强度的对称加密，既能保障信息安全，又能有效降低硬件资源和运算要求。未来可广泛应用于信息安全、数据存储和传输领域。

参考文献：

[1] C. E. Shannon. A mathematical theory of communication. Bell Syst. Tech. J., 27:379–423,623–656, 1948.

[2] T.M.Cover and J.A.Thomas,Elements of Information Theory.New YorK,Wiley 1991.

[3]王平. LZW 无损压缩算法的实现与研究[J]. 计算机工程, 2002, 28(007):98-99.

[4] Huffman D A . A Method for the Construction of Minimum-Redundancy Codes[J]. Resonance, 2006, 11(2):91-99.

[5] G. N. N. Martin, Range encoding: an algorithm for removing redundancy from a digitised message. Video & Data Recording Conference, held in Southampton July 24-27 1979.

[6] Ian H.Witten, Radford M.Neal,John G.Cleary. Arithmetic Coding for Data Compression.Communications of the ACM. 1987,30(6):520~539.

[7] Yang Y， Bamler R， Mandt S . Variable-Bitrate Neural Compression via Bayesian Arithmetic Coding[J]. 2020.

[8] Elminaam D， Kader H， Hadhoud M M . Evaluating the Performance of Symmetric Encryption Algorithms[J]. International Journal of Network Security, 2010, 10(3):213-219.

[9] Biham E， Biryukov A . An Improvement of Davies' Attack on DES[J]. Journal of Cryptology, 1997, 10(3):195-205.

[10] Keller, Nathan, Shamir, et al. Improved Single-Key Attacks on 8-Round AES-192 and AES-256[J]. Journal of cryptology: the journal of the International Association for Cryptologic Research, 2015.

[11] Tromer E， Osvik D A， Shamir A . Efficient Cache Attacks on AES, and Countermeasures[J]. Journal of Cryptology, 2010, 23(1):37-71.

[12]谢琦, 曹雪芹. Blowfish 算法优化及其在 WSN 节点上的实现[J]. 计算机应用与软件, 2013, 30(007):318-320.

[13] Vaucouleurs G， Vaucouleurs A， Corwin J R . The Second Reference Catalogue of Bright Galaxies (RC2). 1976.

[14]万卫星, 李厚强. 可伸缩视频编码的自适应 QP 级联算法[J]. 计算机工程, 2010, 36(006):215-217.

[15]郭媛, 敬世伟. 基于 L-L 级联混沌与矢量分解的无损压缩光学图像加密. 光子学报.

[16] Ho S W， Lai L， Grant A . On the separation of encryption and compression in secure distributed source coding. IEEE.

[17] Mohamed N N， Hashim H， Yussoff Y M , et al. Compression and encryption technique on securing TFTP packet[C]// ISCAIE 2014 - 2014 IEEE Symposium on Computer Applications and Industrial Electronics. IEEE, 2015.

[18]谢冬青, 谢志坚, 李超,等. 关于一种算术编码数据加密方案的密码分析[J]. 通信学报, 2001, 22(003):40-45.

[19]张玉书. 数据压缩与加密的协调性研究[D]. 重庆大学.

[20]WANG Xiao-long, ZHAO Shu-xu. 基于分段线性混沌映射的算术编码与加密[J]. 计算机应用研究, 2014, 31(005):1481-1483.

[21]刘轩黄. 关于随机过程一阶概率分布函数的遍历性[J]. 大学数学, 1989.

[22] Shahid Z， Chaumont M， Puech W . Fast Protection of H.264/AVC by Selective Encryption of CAVLC and CABAC for I and P Frames[J]. IEEE Transactions on Circuits & Systems for Video Technology, 2011, 21(5):565-576.

[23] Xu D . Data hiding in partially encrypted HEVC video[J]. Etri Journal, 2020, 42(3).

[24]贾铸. 算术编码方法在图像压缩编码中的应用[J]. 电子技术:上海, 1999.

作者简介：王杰林，男，1985 年生，湖南平江人，湖南涉外经济学院特聘教授，主要研究方向为编码技术。

廖亦凡，男，1976 年生，湖南长沙人，湖南涉外经济学院信息与机电工程学院副教授，主要研究方向为电路与系统。