doi:10.11918/j.issn.0367-6234.2016.05.015

多元符号的安全算术编码

赵旦峰, 李 超, 薛 睿, 王 杨

(哈尔滨工程大学 信息与通信工程学院, 150001 哈尔滨)

摘 要: 为提高传统算术编码 TAC(traditional arithmetic coding)对密文和选择性明文攻击的抵抗力, 提出一种基于多元符号的安全算术编码 M-SAC(M-ary security arithmetic coding)方案. 该方案将数据压缩与加密相结合, 利用加密密钥改变信源符号在编码区间中的位置, 进而改变 TAC 的编码区间和编码码字. 仿真结果表明: 在静态和自适应模型下, 当译码器利用错误密钥译码时, 该方案比二元随机算术编码 BRAC(binary random arithmetic coding)得到的误符号率 SER(symbol error rate)更高; 该方案既不影响压缩效率又能实现对数据的有效加密, 对密文和选择性明文攻击有很强的抵抗力.

关键词: 数据加密; 数据压缩; 熵编码; 安全算术编码

中图分类号: TN918.91 **文献标志码**: A **文章编号**: 0367-6234(2016)05-0095-05

M-ary secure arithmetic coding

ZHAO Danfeng, LI Chao, XUE Rui, WANG Yang

(College of Information and Communication Engineering, Harbin Engineering University, 150001 Harbin, China)

Abstract: To improve the TAC resistance to ciphertext and selective plaintext attacks, a novel security arithmetic coding scheme based on M-ary symbol is proposed. It combines data encryption with compression. The modified arithmetic coding methodology changes the encoding interval and codeword of TAC by using an encryption key to alter the symbols orders in the encoding interval. Simulations both on static and adaptive model show that M-SAC can obtain higher SER than BRAC when decoding with a wrong key. The proposed encryption algorithm satisfies highly security without loss of compression efficiency, which has strong resistant to ciphertext and selective plaintext attacks.

Keywords: data encryption; data compression; entropy coding; security arithmetic coding

算术编码 AC(arithmetic coding)是一种压缩效率比 Huffman 编码更高的熵编码, 因其良好的压缩性能成为无损压缩的主流, 被广泛应用于各种多媒体压缩标准中, 例如 H.264、JPEG2000 等. 随着 AC 理论研究的逐渐深入, 其安全性也越来越受到诸多学者的关注. 文献[1]提出自适应算术编码具有较好的加密效果, 因其在编码过程中产生了相对随机的编码区间. 文献[2]分析指出自适应算术编码不能抵抗选择性明文攻击. 因此, 为提高传统算术编码 TAC(traditional arithmetic coding)的安全性, 有必要对其编码方法进行修改.

文献[3]最早提出了二元随机算术编码 BRAC(binary random arithmetic coding)的思想. 文献[4]将 BRAC 与混沌系统相结合实现加密, 但损失 6%左右的压缩效率. 文献[5-8]在 BRAC 过程中引入混沌映射模型, 利用混沌映射生成加密密钥, 增加密钥的破译难度. 但从文献[5]可看出, 当用错误密钥译码时, BRAC 得到的误符号率 SER(symbol error rate)不超过 50%. 换言之, 当用错误密钥解密截获的密文时, 能获取明文中一半以上的信息, 所以 BRAC 对密文攻击抵抗力较弱. 文献[9-11]提出区间分割算术编码, 该方法拥有较好的加密效果, 但其实现复杂度高, 并且牺牲了部分压缩效率. 文献[12]提出的安全算术编码采用一阶 Markov 信源模型, 但其在编码过程中需要保存不同条件概率下的编码区间, 将占用较多存储空间.

为进一步提高 TAC 的安全性, 提出一种基于多元符号的安全算术编码 M-SAC(M-ary security arithmetic coding). M-SAC 在编码过程中利用密钥对多元符号在编码区间中的位置进行循环移位, 增

收稿日期: 2014-09-10.
基金项目: 武器装备预先研究(xxxx305030201).
作者简介: 赵旦峰(1960—), 男, 教授, 博士生导师.
通信作者: 李 超, lichao0139@163.com.

加了编码区间分割的随机性.

1 传统算术编码

算术编码的概念最早由 Elias 提出，其核心思想是：从整个符号序列出发，按照符号输入的顺序将符号的概率进行累积，最终把符号序列映射为[0, 1)区间内的一个点. 由于在编码过程中引入了小数，AC 具有很高的压缩效率. 文献[13-14]对其理论研究，AC 已成为一种实际可行的熵编码. 相互独立的概率统计模块和编码模块组成的 AC 比其它熵编码更加灵活.

概率统计模块包括两种模型：静态模型和自适应模型. 静态模型是指在编码过程中符号概率一直不变；自适应模型是指符号概率随着编码过程而改变. 为简化分析，本文讨论静态模型下 AC 的编码原理. 设信源由 3 个符号{A, B, C}组成，其概率和范围见表 1.

表 1 信源符号概率和范围

信源符号	概率	范围
A	0.2	[0, 0.2)
B	0.3	[0.2, 0.5)
C	0.5	[0.5, 1)

从编码区间的下界到上界，将信源符号出现的顺序称之为信源符号在编码区间中的位置. 设输入符号序列为“BACBC”，信源符号在编码区间中的位置为“ABC”，静态 TAC 编码过程见图 1. 编码的初始区间为[0, 1). 当输入第一个符号“B”过后，编码区间缩小为“B”对应的范围[0.2, 0.5). 同时，将区间[0.2, 0.5)按照符号在编码区间中的位置和概率进行分割，为编码下一个符号做准备. 编码下一个符号时，区间进一步缩小为其对应的范围，再进行分割，如此循环，直到编码完最后一个符号. 从图 1 可看出，对符号序列“BACBC”进行编码最终得到的编码区间为[0.2405, 0.245)，在该区间内任取一点可作为编码码字.

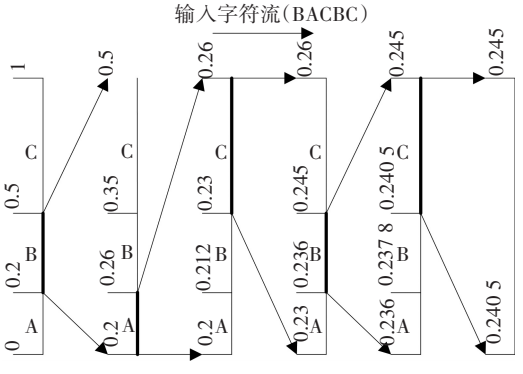

图 1 静态 TAC 编码过程

译码时，根据码字所在区间对应的信源符号来进行译码，译码区间的分割与编码相同. 自适应 TAC 与静态 TAC 唯一的区别是其在编码过程中有符号概率的更新.

2 安全算术编码

2.1 编码理论

从图 1 中可看出，在整个编码过程中信源符号在编码区间中的位置均未改变，这使得 TAC 不能抵抗选择性明文攻击[2]. 本文提出的 M-SAC 很好地解决了这个问题. 与 TAC 的本质区别在于：该方案先利用加密向量改变信源符号在编码区间中的位置，再用 TAC 进行编码，这样就将数据压缩和加密结合在一起，译码时完全依赖于密钥才能正确译码.

加密向量是通过种子生成的与输入符号序列长度相等的二进制序列，密钥是加密向量中的其中一位. 当密钥为 1 时，将信源符号在编码区间中的位置进行循环右移(或左移)；当密钥为 0 时，位置不变.

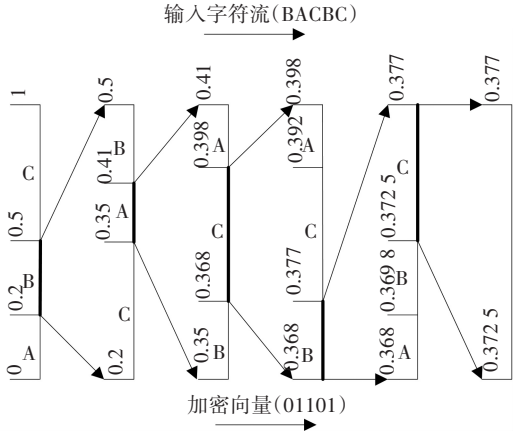

图 2 静态 M-SAC 编码过程

同样以编码符号序列“BACBC”为例. 设信源符号在编码区间中的初始位置为“ABC”，循环移位的方向为右移，加密向量为“01101”.

静态 M-SAC 编码过程见图 2. 经过加密后，信源符号在编码区间中的位置发生了变化. 编码结束时，得到的编码区间为[0.372 5, 0.377)，与图 1 中 TAC 得到的编码区间不同，因此得到的编码码字也不同. 若想正确译码，必须知道加密向量，否则译码失败.

$\boldsymbol{K}$ 为加密向量，k_i 为第 i 个加密密钥，在静态模型下，M-SAC 编/译码步骤为：

1) 用种子生成加密向量 $\boldsymbol{K}$；

2) 令 $i = 0$，初始化信源符号在编/译码区间中的位置，设置循环移位的方向；

3) 如果 $k_i = 1$，先将信源符号在编/译码区间中的位置进行循环移位，再用 TAC 进行编/译码；

如果 $k_i = 0$，直接用 TAC 进行编/译码；

4）i 增加 1，更新 k_i；

5）返回到 3），直到编/译码结束.

在自适应模型下，只需在第 3、4 步之间更新信源符号的概率.

2.2　压缩效率

信息论[15]指出对任何一个独立无记忆信源进行编码，编码一个符号所需的平均比特数不低于该信源的香农熵. 对于三元信源，设 p_A、p_B 和 p_C 分别为信源符号“A”、“B”和“C”的概率，则香农熵为

$$H = -(p_A\log_2 p_A + p_B\log_2 p_B + p_C\log_2 p_C). \tag{1}$$

设 $S = (s_1, s_2, \cdots, s_N)$ 表示长度为 N 的符号序列. 根据 AC 理论，编码这 N 个符号需要的比特数为 $l = -\log_2(p)$，其中 $p = \prod_{i=1}^{N} p(s_i)$，则

$$l = -\sum_{i=1}^{N}\log_2 p(s_i). \tag{2}$$

设编码这 N 个符号所需的实际比特数超过理论值的上限为 ε，于是实际编码每个符号需要的平均比特数 H_S 应满足：$H_S \leqslant [(\varepsilon + l)/N]$，结合式(2)可得

$$H_S \leqslant \frac{\varepsilon - \sum_{i=1}^{N}\log_2 p(s_i)}{N}. \tag{3}$$

设符号序列中“A”的个数为 N_A，“B”的个数为 N_B，“C”的个数为 N_C，即满足 $N = N_A + N_B + N_C$. 当 $N\infty$ 时有：$p_A = N_A/N$，$p_B = N_B/N$，$p_C = N_C/N$. 编码每个符号所需比特数的期望值为

$$\overline{H}_S = E(H_S) \leqslant \frac{\varepsilon - \sum_{i=1}^{N} E[\log_2 p(s_i)]}{N}. \tag{4}$$

其中

$$\sum_{i=1}^{N} E[\log_2 p(s_i)] = N_A\log_2 p_A + N_B\log_2 p_B + N_C\log_2 p_C. \tag{5}$$

式(4) 可化为

$$\overline{H}_S \leqslant \frac{\varepsilon}{N} - \left(\frac{N_A}{N}\log_2 p_A + \frac{N_B}{N}\log_2 p_B + \frac{N_C}{N}\log_2 p_C\right) = \frac{\varepsilon}{N} + H. \tag{6}$$

又编码每个符号所需的比特数不小于香农熵，则 $\overline{H}_S$ 满足

$$H \leqslant \overline{H}_S \leqslant \frac{\varepsilon}{N} + H. \tag{7}$$

所以当 $N \to \infty$ 时，有 $\lim_{N\to\infty} \overline{H}_S = H$.

推导可知：当 N 足够大时，M-SAC 编码一个符号所需的比特数等于该信源的香农熵，能获得最优的压缩效率. 同时，对比图 1、2 的编码过程可发现，最终得到的编码区间不同，但编码区间的长度是相等的，因此 M-SAC 对 TAC 的压缩效率没有影响.

2.3　安全性

M-SAC 利用加密向量改变信源符号在编码区间中的位置，使编码区间分割与密钥相关联，提高了 TAC 对选择性明文攻击的抵抗力. 同时，编码前设定的信源符号在编码区间中的初始位置和循环移位的方向也是一种加密. 因此，对于密文攻击，若不知道信源符号在编码区间中的初始位置、循环移位的方向和加密向量，将无法正确译码. 因此 M-SAC 能有效抵抗密文和选择性明文攻击，安全性非常高.

3　仿真分析

仿真中采用独立无记忆的三符号信源，输入符号序列的长度为 10 000，信源符号在编码区间中的初始位置为“ABC”，循环移位的方向为右移.

3.1　压缩效率

选择 8 种信源，在静态和自适应模型下比较 TAC 和M-SAC的压缩效率，仿真结果见表 2. 其中 H 为信源的香农熵；E 为熵编码界（编码输入符号序列所需比特数的下限），$E = N \times H = 10\ 000$ H；T_{AC} 为用 TAC 输出的比特数；S_{AC} 为用 M-SAC 输出的比特数；η 为M-SAC输出的二进制序列中 1 与 0 的个数比；λ 为 M-SAC 输出序列相对于 TAC 输出序列的变化率.

从表 2 中可看出，当信源香农熵和概率模型一定时，M-SAC 与 TAC 输出的比特数相等或相差 1. 这是因为 AC 具体实现时精度有限，在编码过程中改变信源符号在编码区间中的位置，会导致输出的比特数与 TAC 相比有 1 bit 左右的差值. 由此可得，M-SAC 对 TAC 的压缩效率没有影响. 而文献[4,9]中提出的安全算术编码方案损失了部分压缩效率，文献[4]中损失的压缩效率达到 6%.

同时在两种概率模型下，M-SAC输出的二进制序列中 1 与 0 的个数比都在 1 左右，说明密文中“1”和“0”的个数相对平衡，达到了理想的加密效果. 从 λ 值可知，M-SAC 输出序列相对于 TAC 输出序列的变化率在 50%左右，说明经过加密编码输出序列变化随机性强，安全性高.

3.2　密钥敏感性

对于二进制密文，如果改变加密向量中某一位或某几位，密文的变化率达到 50%，说明加密向量敏感性强.

表 2 TAC 与 M-SAC 的压缩效率

H/ bit/symbol	E/ bit	静态模型				自适应模型			
		T_{AC}/bit	S_{AC}/bit	η	λ/%	T_{AC}/bit	S_{AC}/bit	η	λ/%
0. 090 8	908	909	909	0. 942 3	49. 73	926	926	0. 852 0	47. 52
0. 306 1	3 061	3 062	3 061	1. 061 3	48. 66	3 077	3 077	0. 937 7	50. 24
0. 509 4	5 094	5 095	5 095	1. 001 2	50. 93	5 109	5 109	1. 029 8	50. 17
0. 703 9	7 039	7 040	7 040	1. 017 8	49. 38	7 053	7 053	1. 012 8	50. 06
0. 905 6	9 056	9 058	9 057	1. 005 5	50. 62	9 070	9 070	0. 982 5	50. 25
1. 103 7	11 037	11 038	11 037	1. 000 1	49. 98	11 051	11 050	1. 034 2	49. 74
1. 302 1	13 021	13 023	13 022	0. 973 0	50. 78	13 035	13 034	0. 990 2	50. 82
1. 509 2	15 092	15 093	15 093	0. 977 3	50. 81	15 106	15 106	1. 004 3	49. 86

图 3 为当信源香农熵为 1. 302 1 bit/symbol 时，改变加密向量，在不同种子和概率模型下 M-SAC 得到的密文的变化率. 可看出，对于不同的种子和概率模型，无论改变多少个加密密钥，密文的变化率都在 50%左右，所以加密向量具有很强的敏感性.

表 3 为在两种概率模型下，当信源的香农熵为 1. 302 1 bit/symbol (p_A = 0. 105 0、p_B = 0. 294 0)时，解密向量的敏感性. 其中 X 为改变解密密钥的个数，H' 和 e 分别为译码符号序列的香农熵和 SER，p'_A 和 p'_B 分别为译码符号序列中“A” 和“B” 的概率.

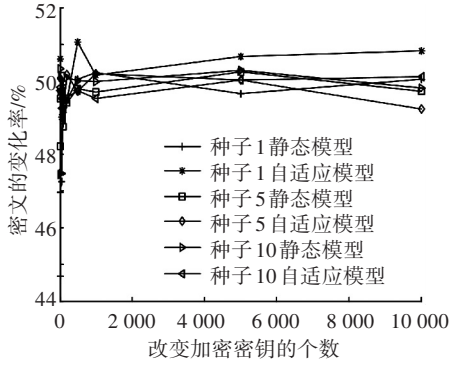

图 3 改变加密向量密文的变化率

表 3 解密向量的敏感性

X	静态模型				自适应模型			
	p'_A	p'_B	H'/bit · symbol^{-1}	e/%	p'_A	p'_B	H'/bit · symbol^{-1}	e/%
5	0. 100 3	0. 292 2	1. 288 2	53. 36	0. 322 1	0. 571 9	1. 331 7	78. 95
10	0. 102 6	0. 291 2	1. 293 1	53. 51	0. 319 9	0. 562 8	1. 355 4	75. 05
50	0. 100 8	0. 291 3	1. 288 6	53. 86	0. 385 6	0. 533 3	1. 307 7	75. 71
100	0. 108 7	0. 291 3	1. 308 5	54. 79	0. 327 7	0. 516 0	1. 438 5	71. 97
200	0. 101 2	0. 289 1	1. 287 3	53. 79	0. 367 1	0. 516 6	1. 384 0	74. 00
500	0. 103 3	0. 296 4	1. 300 3	54. 13	0. 149 1	0. 674 4	1. 234 3	67. 98
1 000	0. 107 7	0. 291 1	1. 294 0	53. 61	0. 484 9	0. 153 5	1. 452 0	68. 19
5 000	0. 100 4	0. 297 5	1. 294 0	53. 61	0. 576 4	0. 314 6	1. 331 6	78. 28
10 000	0. 109 2	0. 283 4	1. 301 3	54. 22	0. 218 3	0. 576 1	1. 406 9	68. 06

从表 3 可看出，在两种概率模型下，改变解密向量，译码符号序列的 p'_A、p'_B 和 H' 与输入符号序列相比都发生了改变，SER 超过 50%. 在自适应模型下，SER 甚至接近 80%. 改变解密密钥的个数对 SER 没有太大影响，说明解密向量敏感性强.

为比较两种概率模型的加密安全性，表 4 给出了在静态模型和自适应模型下，针对不同的信源，运用错误密钥(与加密向量有 10 位不同)解码，得到的译码序列的香农熵(H') 和 SER(e)，及译码序列中符号“A” 和“B” 的概率p'_A 和p'_B. 其中p_A 和p_B 分别为输入序列中符号“A” 和“B” 的概率.

表 4 不同信源的加密安全性

H/bit · symbol^{-1}	p_A	p_B	静态模型				自适应模型			
			p'_A	p'_B	H'/bit · symbol^{-1}	e/%	p'_A	p'_B	H'/bit · symbol^{-1}	e/%
0. 090 8	0. 005 0	0. 005 0	0. 004 9	0. 005 7	0. 095 3	2. 04	0. 005 1	0. 992 0	0. 074 0	99. 30
0. 306 1	0. 015 0	0. 030 0	0. 014 8	0. 031 6	0. 312 8	8. 71	0. 023 5	0. 965 0	0. 250 8	96. 41
0. 509 4	0. 025 0	0. 064 0	0. 025 4	0. 066 6	0. 521 3	16. 43	0. 068 2	0. 905 4	0. 532 4	92. 41
0. 703 9	0. 035 0	0. 107 0	0. 034 9	0. 106 7	0. 702 5	24. 92	0. 102 0	0. 866 5	0. 672 2	88. 63
0. 905 6	0. 045 0	0. 168 0	0. 043 4	0. 165 7	0. 893 8	35. 06	0. 147 3	0. 820 0	0. 803 1	84. 74
1. 103 7	0. 055 0	0. 256 0	0. 055 3	0. 260 1	1. 110 6	45. 95	0. 249 4	0. 692 4	1. 105 6	79. 35
1. 302 1	0. 105 0	0. 294 0	0. 102 6	0. 291 2	1. 293 1	53. 51	0. 319 9	0. 562 8	1. 355 4	75. 05
1. 509 2	0. 205 0	0. 328 0	0. 207 4	0. 330 8	1. 513 4	63. 49	0. 390 2	0. 441 1	1. 483 8	72. 70

表 4 表明在静态模型下译码序列中符号“A” 和“B” 的概率与输入序列中对应符号的概率非常相近；而在自适应模型下，译码序列中符号的概率与输入序列中对应符号的概率相比发生了很大变化. 同时，在静态模型下 SER 随信源熵增加而增大，但都小于自适应模型下的 SER. 另外，自适应模型下 H' 的改变程度比在静态模型下更大. 由此可知，自适应模型比静态模型加密安全性更高，更难破译.

这是因为在自适应模型下符号的概率随编码过程不断地更新；而在静态模型下，符号的概率在编码过程中一直不变.

当用错误密钥译码时，与文献[5]中的 BRAC 相比，M-SAC 得到的 SER 更高，尤其是在自适应模型下. 故 M-SAC 的加密效果更好，对密文的抵抗力更强.

3.3　明文敏感性

明文敏感性是指改变明文中的某一位或某几位，密文的变化情况. 当信源熵为 1.302 1 bit · symbol^{-1}时，改变明文，在不同种子和概率模型下 M-SAC 得到的密文的变化率见图 4.

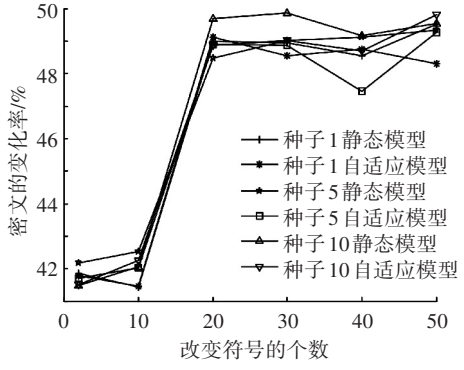

图 4　改变明文密文的变化率

图 4 表明对于不同的种子和概率模型，当改变明文中 2～50 个符号时，密文的变化率在 41%～50%. 也就是说，明文稍微改变就会引起密文的随机变化，说明明文敏感性较强.

4　结　论

本文提出基于多元符号的安全算术编码，将数据压缩和加密结合在一起. 利用密钥对信源符号在编码区间中的位置进行循环移位，改变 AC 的编码区间，进而改变编码码字. 理论分析 M-SAC 的压缩效率和安全性. 仿真表明，当利用错误密钥译码时，在自适应模型下，M-SAC 能得到 70%以上的 BER，比 BRAC 加密效果更好. 同时，M-SAC 对压缩效率没有影响，并且具有很强的密钥敏感性和明文敏感性，对密文和选择性明文攻击有很强的抵抗力. 讨论三元符号的安全算术编码，更高元的安全算术编码方法与其类似. 随着编码元数的增加，M-SAC 的安全性会更高，但其复杂度也会上升. 在实际应用中，平衡加密安全性和复杂度，选择合适的编码元数. 该方案可应用于对密文攻击抵抗力要求高的系统中.

参考文献

[1] WITTEN I H, CLEARY J G. On the privacy afforded by adaptive text compression [J]. Computer Security, 1988, 7: 397-408.

[2] BERGEN H A, HOGAN J M. A chosen plaintext attack on an adaptive arithmetic coding compression algorithm [J]. Computer Security, 1993, 12: 157-167.

[3] GRANGETTO M, MAGLI E, OLMO G. Multimedia selective encryption by means of randomized arithmetic coding [J]. IEEE Transactions on Multimedia, 2006, 8(5): 905-917.

[4] BOSE R, PATHAK S. A novel compression and encryption scheme using variable model arithmetic coding and coupled chaotic system [J]. IEEE Transactions on Circuits and Systems, 2006, 53(4):848-857.

[5] LI Hengjian, ZHANG Jiashu. A secure and efficient entropy coding based on arithmetic coding [J]. Communications on Nonlinear Science and Numerical Simulation, 2009, 14: 4304-4318.

[6] WONG K, LIN Qiuzhen, CHEN Jianyong. Simultaneous arithmetic coding and encryption using chaotic maps [J]. IEEE Transactions on Circuits and Systems, 2010, 57(2): 146-150.

[7] 代才莉，包万宇. Logistic 映射控制的安全算术编码及其在图像加密中的应用[J]. 重庆大学学报，2012，35(8):87-91.

[8] 王小龙，赵庶旭. 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(5)：1481-1483.

[9] WEN Jiangtao, KIM H, VILLASENOR J D. Binary arithmetic coding with key-based interval splitting [J]. IEEE Signal Processing Letters, 2006, 13(2): 69-72.

[10] KIM H, WEN Jiangtao, VILLASENOR J D. Secure arithmetic coding [J]. IEEE Transactions on Signal Processing, 2007, 55(5): 2263-2272.

[11] ZHOU Jiantao, AU O C, WONG P H. Adaptive chosen-ciphertext attack on secure arithmetic coding [J]. IEEE Transactions on Signal Processing, 2009, 57(5): 1825-1838.

[12] DUAN Lili, LIAO Xiaofeng, XIANG Tao. A secure arithmetic coding based on Markov model [J]. Communications on Nonlinear Science and Numerical Simulation, 2011, 16: 2554-2562.

[13] RISSANEN J, LANGDON G G. Arithmetic Coding [J]. IBM Journal of Research and Development, 1979, 23(2): 149-162.

[14] WITTEN I H, NEAL R M, CLEARY J G. Arithmetic coding for data compression [J]. Communications of the ACM, 1987, 30(6): 520-540.

[15] SHANNON C E. A mathematical theory of communication [J]. The Bell System Technical Journal, 1948, 27: 379-423, 623-656.

（编辑　王小唯　苗秀芝）