엔트로피 방법을 이용한 평문·암호문 식별방법에 관한 연구

차경준* , 류제선*

* 한양대학교 수학과

On discernment of plain and cipher text using the entropy test

Kyung Joon Cha*, Je Seon Ryu**

* Department of Mathematics, Hanyang Univ.

요 약

암호 알고리즘 출력문에 대한 난수성 검정들은 평문과 암호문 식별에 중요한 역할을 하고 있다. 실제로, 난수열의 생성자는 비밀키의 생성자와 같은 많은 암호체계에서 사용되고 있으며, 이때 사용되고 있는 난수열은 모의 난수라고 한다. 따라서, 이진수열에 대한 난수성을 검정하는 통계적 검정방법이나 다른 이론적 기준이 필요하다. 본 논문에서는 모의난수열이 갖고 있는 난수성 판정에 관하여 universal 엔트로피 검정방법과 근사 엔트로피 검정방법을 이용하며, 위의 두 방법에 대한 각각의 이론적인 배경과 모의실험을 통한 판정기준을 제공한다.

I. 서 론

최근의 암호화 기법의 응용에 있어 난수열 혹은 모의난수열에 대한 요구가 급증되고 있으며 주어진 수열의 난수성 판단에 근거하여 암호 알고리즘의 안전성을 평가하고 있는 실정에 있다. 즉, 암호 알고리즘 설계시 출력된 암호문이 난수성을 만족할 때 좋은 암호 알고리즘으로 판정되고 있으며 난수성을 검정하기 위해 암호 알고리즘을 통해 출력된 결과의 통계적인 성질을 살펴봄으로서 판정할 수 있다.

기존에 사용된 통계적 검정방법으로는 frequency test, poker test, runs distribution test, serial test 등이 있으며, 박홍구 등과 한국정보보호센타는 이러한 통계적 방법들을 이용하여 검정을 위한 최소한의 표본크기에 대한 연구를 하였다. [1][2][3][4][5][6][7][8][9]

엔트로피 관점에서 이진수열에 대한 난수성을 고려해 보면, 암호학적으로 중요한 정량화된 측도인 비트당 엔트로피를 추정하여 수열의 난수성에 대하여 검정하는 데에 있다. 엔트로피 검정방법은 기존의 통계적 관점의 검정에 비해 좀더 일반적인 통계적 모형을 기반으로 하며, 공격자가 비밀키 소스의 통계적 결점에 관한 지식을 부당하게 이용할 때 공격자가 수행하는 최적키 탐색 전략의 시간과 관련이 있는 소스의 비트당 엔트로피를 측정한다. 또한, 다른 길이의 인접한 부분수열의 빈도를 이용하여 주어진 수열의 길이와 부분수열의 길이에 대하여 엔트로피 값을 추정함으로서 주어진 수열의 추정된 엔트로피의 편차를 측정한다. 첫번째 방법이 universal 엔트로피 검정방법이며, 두번째 방법이 근사 엔트로피 검정방법이다. [10][11][12]

본 연구에서는 보편적으로 많이 사용되고 있는 엔트로피 정량화 방법인 universal 검정방법과 근사 엔트로피 검정방법을 주어진 암호문과 평문 이진수열에 적용하여, 주어진 유의수준에서 암호문이 암호문으로, 평문이 평문으로 판단되는 성공률을 비교, 분석한다. 평문으로서 5개의 임의의 문서파일(cpp, doc, pdf, ps, tex)을 선정하고, 임의의 키와 위에 주어진 평문들을 이용하여 DES(Data Encryption Standard)[13][14]를 CBC(Cipher Block chaining) 모드로 생성한 수열을 암호문으로 가정한다. 이러한 이진수열에 대해서 평문 및 암호문 식별에 대한 엔트로피 정량화 방법의 특성과 성공률을 분석한다.

II. 분석 및 비교 방법

1. universal 엔트로피 검정방법

독립이며 확률이 1/2인 N개의 베르누이 시행으로 주어진 이진수열에서 길이가 L인 서로 다른 블록이 발생할 확률은 $1/2^L$이다. 즉, 같은 pattern의 블록이 발생하는 기대 거리(expected distance)는 2^L이다. 따라서, $\log_2$[기대 거리]는 블록의 길이 L이며, 이것이 주어진 이진수열에 대한 N 비트 당 엔트로피, 즉 평균 정보량이다. 검정 통계량 T는 양의 정수 값인 파라미터 L, Q, K에 의해 설정된다. 검정을 수행하기 위해, 먼저 $s^N = s_0 s_1 s_2 \cdots s_{N-1}$을 전체의 길이가 N인 이진수열이라 하자. 먼저 s^N에서 길이가 L인 N/L개의 겹치지 않는 블록을 만든다. 처음 Q개의 블록을 초기화를 위한 블록으로 사용하고 나머지 $K = N/L - Q$개의 블록은 검정을 위해 사용된다. 검정을 위한 n번째 블록을

$$b_n(s^N) = [s_{Ln}, s_{Ln+1}, \cdots, s_{L(n+1)-1}]$$

으로 정의하고 $Q \le n \le Q+K-1$에 대하여 $A_n(s^N)$을 n번째 블록 $b_n(s^N)$과 처음으로 일치하게 되는 $b_{n-i}(s^N)$가 존재할 때의 i로 정의하며, 존재하지 않는 경우 n으로 정의한다. 얻어진 결과 값을 이용하여 다음의 통계량 T를 계산한다.

$$T(s^N) = \frac{1}{K}\sum_{n=Q}^{Q+K-1} \log_2 A_n(s^N) \qquad (1)$$

$Q \to \infty$ 일 때, $T(s^N)$의 기대값과 분산은

$$E(T(s^N)) = \sum_{j=1}^{\infty} 2^{-L}(1-2^{-L})^{j-1}\log_2 j \qquad (2)$$

$$Var(T(s^N)) = \sum_{j=1}^{\infty} 2^{-L}(1-2^{-L})^{j-1}(\log_2 j)^2 - E(T(s^N))^2 \qquad (3)$$

이다[10]. 따라서, $T(s^N)$의 분산은 서로 독립인 이진수열에서 $K \to \infty$ 일 때

$$\lim_{K\to\infty} Var(T(s^N)) = \frac{Var(\log_2 A_n(s^N))}{K}$$

표 1 주어진 블록의 길이에 따른 $T(s^N)$의 평균과 분산

L	$E(T(s^N))$	$Var(T(s^N))$
6	5.2177052499	2.9540323994
7	6.1962506541	3.1253918686
8	7.1836655535	3.2386621610
9	8.1764247579	3.3112008795
10	9.1723243082	3.3564569070
11	10.170032292	3.3840870307
12	11.168764874	3.4006541451
13	12.168070314	3.4104380092
14	13.167692567	3.4161418218
15	14.167488449	3.4194303978
16	15.167378764	3.4213083430

이다. 따라서, Eq.(2),(3)을 이용하여 귀무가설 H_0을 검정하기 위해 검정 통계량

$$z = \frac{T(s^N) - E(T(s^N))}{\sqrt{(Var(T(s^N))}} \qquad (4)$$

을 얻을 수 있고 양측검정을 시행할 수 있다. $6 \le L \le 16$, $Q \ge 10 \cdot 2^L$, 그리고 $K \ge 1000 \cdot 2^L$에 대하여 $E(T(s^N))$과 $Var(T(s^N))$이 <표 1>에 있다.

2. 근사 엔트로피 검정방법

Pincus[11]와 Pincus와 Singer[12]는 근사 엔트로피를 이용하여 불규칙성의 정도에 관하여 일련의 자료에서의 난수성의 척도의 개념을 공식화하였다. 근사 엔트로피의 개념은 다음과 같다.

양의 정수 N과 $m \le N$인 비음 정수 m이 주어졌을 때 양의 실수 r과 실수열 u: $(u(1), u(2), \cdots, u(N))$에 관하여 m블록 $x(i)$와 $x(j)$ 사이의 거리를 $d(x(i), x(j))$으로 정의하고, 또한,

$$Q_i^m(r) = \frac{d(x(i), x(j) \le r\text{인 } j \le N-m+1\text{의 개수})}{N-m+1}$$

라고 하면, $Q_i^m(r)$은 주어진 블록과 근사적으로 같은 블록들의 상대적 빈도라고 할 수 있다.

Pincus와 Singer[12]는 수열 u의 근사 엔트로

피를 측정하는 계산적 방법으로서

$$ApEn(m,r,N)(u)=\phi^m(r)-\phi^{m+1}(r),\ m\geq 1$$

을 제시하였으며, 여기서

$$\phi^m(r)=\frac{1}{N-m+1}\sum_{i=1}^{N-m+1}\log_2 Q_i^m(r).$$

0과 1을 포함하는 이진수열에서, $r<1$이라 할 수 있으며, 두 m블록의 거리 $d(x(i),x(j))$는 0과 1의 값만을 갖는다. 따라서, $r=0$으로 제한할 수 있으며, $ApEn$은 r과 독립적으로 사용될 수 있다.

성공률이 p인 p랜덤 베르누이 수열에서 $\psi^m(p)$를 ϕ^m의 기대 엔트로피라고 하면

$$\psi^m(p)=\sum_{t=0}^{m}\binom{m}{t}p^{m-t}(1-p)^t\log p^{m-t}(1-p)^t \tag{5}$$

으로 정의된다. 이때, $T_1(p,m)$을 (추정된) 엔트로피 ϕ^{m+1}과 Eq.(5)에서 계산된 기대 엔트로피 ψ^{m+1}의 평균편차라고 정의하자. 즉,

$$T_1(p,m)=\frac{\phi^{m+1}-\psi^{m+1}(p)}{m+1} \tag{6}$$

일 때, $T_1(p,m)$은 p랜덤 베르누이 과정에 의해 생성된 유한 수열에 대하여 0에 가까운 값을 갖게 된다. 즉, $T_1(p,m)$ 값이 작으면 랜덤한 수열로, 큰 값을 가지면 랜덤하지 않은 수열로 해석할 수 있다. Eq.(6)에서 $p=0.5$인 베르누이 수열의 모든 m에 대하여 $\psi^m(p)=-m\log 2$이므로

$$T_1(0.5,m)=\frac{\phi^{m+1}+m\log 2}{m+1} \tag{7}$$

이다. 따라서, Eq.(7)에서의 통계량 $T_1(0.5,m)$을 고려했을 때, 주어진 m과 유의수준에서 $T_1(0.5)$값이 커질 때, 귀무가설 H_0을 기각할 수 있다. 따라서, H_0이 참일 때, 관측된 m블록의 빈도는 수열의 길이 N이 증가하면 기대 빈도로 근사한다. 즉, $N\to\infty$일 때, $T_1(p)$는 확률적으로 0에 근사한다. 결국, N이 증가할 때 관측된 엔트로피 ϕ^{m+1}은 기대 엔트로피 ψ^{m+1}에 근사하며, Eq.(7)에서의 분자는 0에 근사한다.

표 2 $p=0.5$에 대하여 10000번의 실험에서 얻어진 유의수준 0.01, 0.05, 그리고 0.10에서의 T_1에 대한 임계값

N \ α	0.01	0.05	0.10
512	0.0106148301	0.0077741937	0.0065182655
1024	0.005380361	0.0038349925	0.0032030667
2048	0.0026892232	0.0019269289	0.0016101718
4096	0.0013132284	0.000961751	0.0008091151

본 연구에서는 <표 2>에서 $T_1(0.5,m)$에 대한 0.01, 0.05, 그리고 0.10의 유의수준에 따른 임계값을 제공하고자 한다. 이러한 문제를 해결하기 위해 $p=0.5$인 베르누이 확률과정을 고려하였고, 실제 이진수열의 랜덤성을 조사하기 위해 요구되어지도록 길이가 $N=2^9, 2^{10}, 2^{11}, 2^{12}$인 각각의 수열을 10000번의 시험 후 생성하였다. 실험에서 구한 임계값을 이용하여 비선형 회귀모형을 최소제곱법으로 추정해 보면 임의의 수열의 길이 N과 $\alpha=0.01,0.05,0.10$에 대하여 각각

$$T_1=0.0170\exp(-0.000964N),\ R^2=0.99,$$

$$T_1=0.0127\ \exp(-0.00102N),\ R^2=0.99$$

$$T_1=0.0106\exp(-0.00103N),\ R^2=0.98$$

으로 추정되었다. 여기서 R^2은 결정계수이다. 그러므로, 임의의 N과 주어진 유의수준에 대하여 위에서 구해진 회귀선을 이용하면, T_1에 대한 임계값을 구할 수 있다.

Ⅲ. 결과 및 분석

본 연구에서는 임의의 5개의 파일에서 생성된 이진수열과, 이를 이용한 암호 알고리즘을 통해 생성된 새로운 이진수열의 난수성을 판별하기 위해 universal 엔트로피 검정방법과 근사 엔트로피 검정방법을 사용하여 두 이진수열의 p-값을 확인하였다.

1. universal 엔트로피 검정 결과

본 연구에서의 universal 엔트로피 검정방법은 $N=409602$인 전체 수열을 선택한 후 주어진 조

표 3 주어진 블록의 길이에 대한 전체 수열의 길이와 저장된 소스의 길이

L	$\geq N$	$Q \geq 10 \cdot 2^L$
6	387840	640
7	904960	1280
8	2068480	2560
9	4654080	5120
10	10342400	10240
11	22753280	20480
12	49643520	40960
13	107560960	81920
14	231669760	163840
15	496435200	327680
16	1059061760	655360

표 4 선택된 파일에 대한 universal 엔트로피 검정에서의 $T(s^N)$, z-값, 그리고 p-값

	평문		
파일	$T(s^N)$	z-값	p-값
cpp	4.916515	-45.52876	0
doc	3.760216	-220.3179	0
pdf	5.185298	-4.898805	4.821e-7
ps	4.356115	-130.2401	0
tex	4.859396	-54.16301	0
파일	암호문		
cpp	5.219149	0.2182621	0.5863876
doc	5.223413	0.8628526	0.8058907
pdf	5.213769	-0.5949632	0.275934
ps	5.219344	0.2477839	0.5978492
tex	5.226411	1.316054	0.9059221

건에 의하여 길이 $L = 6$인 블록을 이용하여 처음 $Q = 767$개의 초기화 블록과 나머지 $K = 67500$개의블록을 검정한 결과이다. 이는 Q 블록 내에서 높은 확률로서 L비트의 패턴이 적어도 한번 나타나게 하기 위하여 Maurer[15]에서 L이 $6 \leq L \leq 16$, $Q \geq 10 \cdot 2^L$이 되도록 제안된 결과를 이용한 것이다. 실제로, 블록의 크기 L이 $6 \leq L \leq 16$일 때, 제안된 Q블록의 개수와 전체

표 5 선택된 파일에 대한 근사 엔트로피 검정에서의 평균 T_1 값과 성공률

	평문		암호문	
파일	평균 T_1	성공률(%)	평균 T_1	성공률(%)
cpp	0.01129083	100	0.000463782	96
doc	0.1081114	100	0.000459994	97
pdf	0.001535578	47	0.000466323	95
ps	0.0538687	100	0.000399024	98
tex	0.01138764	100	0.000460892	98

이진수열의 개수는 <표 3>과 같다.

<표 4>를 보면 평문의 경우, pdf 파일은 p-값이 다른 파일과 비교하여 상당히 큰 값을 갖는 것으로 보아 파일 자체가 갖고 있는 비트당 엔트로피의 성질이 다른 파일들보다는 높기 때문으로 사료된다. 다른 평문 파일들은 0에 가까운 p-값을 갖고 있는 것으로 보인다. 암호문의 경우 모든 파일에서 통계량의 값이 기대값 주위에 모여 있는 것으로 보이며, 따라서 양측검정을 하는 경우 모두 난수성을 만족하고 있는 것으로 판단된다

2. 근사 엔트로피 검정 결과

근사 엔트로피 검정방법은 다른 길이의 인접한 부분수열의 빈도를 이용하여 주어진 수열의 길이와 부분수열의 길이에 대한 엔트로피 값을 측정함으로서 주어진 수열의 추정된 엔트로피의 편차를 측정하는 방법이다. <표 5>는 길이 N이 $N = 4096$인 전체수열 100개를 선택한 후 주어진 조건에 의하여 길이 $m_{crit} = 3$인 블록을 이용하여 검정한 결과이다. 평문의 경우 universal 엔트로피 검정방법과 비슷하게 pdf 파일에서 평균 통계량 값이 비교적 작으며 성공률 또한 상당히 낮았음을 알 수 있다. 이는 주어진 수열의 부분 수열에 대하여 추정된 엔트로피의 편차가 작기 때문에 발생하는 것으로 판단되며, 결국 성공률이 낮게 판정되었다. 또한 pdf 파일을 제외한 모든 파일에서는 높은 통계량 값을 갖고 있으며, 결국 100%의 성공률을 보이고 있다. 암호문의 경우 통계량 값이 상당히 낮고 안정되어 있음을 알 수 있었으며 성공률에 있어서 모든 파일에서 95% 이상으로 안정되었음을 알 수 있다.

Ⅳ. 결 론

이번 연구에서 universal 엔트로피 검정방법과

근사 엔트로피 검정방법에 따른 5개의 파일에 대한 p-값을 구하였다. 이때, 각각의 파일에 대한 실험에서 통계량 값과 p-값의 변화량을 볼 수 있었으며, 전체적으로 평문보다는 암호문에 대한 p-값이 안정적이라는 것을 알 수 있었고, 두 엔트로피 방법의 결과에서 유사한 결과를 얻을 수 있다는 것을 볼 수 있었다. 따라서 평문 판정 알고리즘에서는 이진수열에 내재되어 있는 복잡한 경향성을 검정할 수 있는 다양한 방법을 채택하는 것이 바람직하리라 사료되며 박홍구 등에서 제안된 통계적 방법등과 함께 고려하여 모의 난수에 대한 난수성 검정방법을 보다 다각적인 측면으로 고려하는 것이 좋은 암호 알고리즘을 개발하는데 중요한 역할을 할 것으로 사료된다.[8] 또한, 평문, 암호문 식별 알고리즘 설계시 universal 엔트로피 검정방법과 근사 엔트로피 검정방법을 이용하여 주어진 임계값 조건에 대하여 판정한다면 주어진 유의수준에 의하여 좋은 결과를 얻을 수 있으리라 사료된다.

참 고 문 헌

[1] C.E. Shannon, Communication theory of secrecy systems, Bell System Technical Journal, 28, 656-715, 1949

[2] H. Beker and F. Piper, Cipher Systems: The Protection of Communications, Wiley, 1982

[3] A.M. Mood, The distribution Theory of Runs, Annals of Mathematical Statistics, 11, 367-392, 1940

[4] I.J. Good, The Serial Test for Sampling Numbers and Other Tests for Randomness, Proceedings of the Cambridge Philosophical Society, 49, 276-284, 1953

[5] I.J. Good, On the Serial Test for Random Sequences, Annals of Mathematical Statistics, 28, 262-264, 1957

[6] D.J. Sheskin, Handbook of Parametric and Nonparametric Statistical Procedures, CRC Press Inc, 140-143, 1997

[7] J.D. Gibbons, Nonparametric Statistical Inference, 2nd. edition, Marcel Dekker Inc, NewYork, 68-90, 1985

[8] 박홍구, 차경준, 장호종, 송정환, 박성준, 실험계획법을 이용한 평문, 암호문 식별방법의 검정크기에 관한 연구, 통신정보보호학회논문지, 19, 4, 71-84, 2000

[9] 한국정보보호센타, 암호통신문 탐지기술 연구 및 S/W 개발, 1999

[10] U.M. Maurer, A universal statistical test for random bit generator, Journal of Cryptology, 5, 2, 89-105, 1992

[11] S. Pincus, Approximate entropy as a measure of system complexity, Proceedings of the National Academy of Sciences, 88, 2297-2301, 1991

[12] S. Pincus and B.H. Singer, Randomness and degrees of irregularity, Proceedings of the National Academy of Sciences, 93, 2083-2088

[13] FIPS, Data Encription Standard, Federal Information Processing Standard(FIPS) Publication, 46, National Bureau of Standards, Washington DC., 1977

[14] 현대암호학, 한국전자 통신연구소, 58-66, 1991

[15] S. Chatterjee, M.R. Yilmaz, M. Habibullah and M. Laudato, An approximate entropy test for randomness, Commun. Statist.-Theory Meth., 29, 3, 655-675, 2000