LZ77 압축 알고리즘 기반 최소 엔트로피 추정 기법

강근호, 김지훈, 우지헌, 김용준*
포항공과대학교

{bestkevin63, enzer0, jhwoo1997, yongjune} @postech.ac.kr

On the Min-Entropy Estimation via LZ77 Compression Algorithm

Geunho Kang, Jihun Kim, Jiheon Woo, Yongjune Kim
POSTECH

요 약

암호 및 보안 기술의 중요성이 확대됨에 따라, 더욱 신뢰성 높은 난수의 필요성이 커지고 있다. 난수의 랜덤성을 판별하기 위해 정보이론의 최소 엔트로피가 널리 활용되며, 국제 표준인 NIST SP 800-90B 에서 이를 추정하기 위한 10 가지 평가 방법을 제시하고 있다. 본 연구에서는 LZ77 압축 알고리즘을 사용하여 난수열의 최소 엔트로피를 추정하는 새로운 기법을 제시하고자 한다. LZ77 알고리즘의 압축률이 이론적으로 섀넌 엔트로피에 수렴한다는 정리에 기반하여, 섀넌 엔트로피를 추정한 후 이를 바탕으로 최소 엔트로피를 도출하는 새로운 최소 엔트로피 추정 기법을 제안하고자 한다.

Ⅰ. 서 론

본 논문에서는 LZ77 압축 알고리즘을 활용하여 시퀀스의 최소 엔트로피(min-entropy)를 추정하는 새로운 방법을 제안한다. 암호 시스템의 보안성은 난수의 무작위성에 의존하며, 엔트로피는 시퀀스의 무작위성을 정량화하는 대표적인 측도로 활용된다. 이 중 최소 엔트로피는 다양한 엔트로피의 정의 중 가장 보수적인 하한값을 제공하며, 무작위성 판별에 널리 활용된다.

난수를 생성하는 엔트로피 소스의 성능을 평가하기 위해, 국제 표준인 NIST SP 800-90B 에서는 최소 엔트로피를 추정하기 위한 10 가지 평가 방법을 제시하고 있다 [1]. 본 연구에서는 새로운 추정 기법을 제안하며, 기존 기법과의 비교를 통해 제안 방법의 우수성을 입증하고자 한다.

LZ77 알고리즘의 압축률은 입력 시퀀스의 섀넌 엔트로피에 수렴함이 알려져 있다 [2]. 이에 본 연구는 LZ77 기반 압축률을 통해 시퀀스의 섀넌 엔트로피를 추정하고, Fano's inequality 를 이용하여 최소 엔트로피를 유도하는 방법을 제안한다.

Ⅱ. 이론적 배경

LZ77 압축 알고리즘은 다양한 유형의 시퀀스에 범용적으로 활용되는 압축 기법으로, 시퀀스에 대한 사전 정보 없이 반복되는 통계적 구조를 포착하여 데이터를 압축한다. 이 알고리즘은 압축이 완료된 이전 시퀀스를 저장하는 윈도우 버퍼(window buffer)와, 압축할 시퀀스를 임시로 보관하는 룩어헤드 버퍼(lookahead buffer)를 활용한다.

LZ77 압축 알고리즘의 작동 방식은 다음과 같다. 먼저, 압축할 문자열을 입력으로 받아 룩어헤드 버퍼에 순서대로 채운다. 이후, 룩어헤드 버퍼에 있는 문자열 중 윈도우 버퍼에 존재하는 가장 긴 일치 문자열을 탐색한다. 최장 일치 문자열이 발견되면, 해당 문자열의 상대 위치(distance), 길이(length), 그리고 일치 문자열 바로 뒤에 위치한 다음 문자(character)를 묶어 (D, L, C) 형태의 트리플로 출력한다. 이후 버퍼는 $(L+1)$ 만큼 앞으로 이동하며, 룩어헤드 버퍼의 빈 공간에는 남은 입력 문자열이 채워진다. 이 과정을 반복하여 룩어헤드 버퍼가 모두 소진되면 압축이 완료된다.

LZ77 압축 알고리즘이 산출하는 압축 길이는 입력 시퀀스의 섀넌 엔트로피에 점근적으로 수렴하는 것으로 알려져 있다 [2]. 따라서 본 연구에서는 LZ77 압축률을 섀넌 엔트로피율의 추정값으로 간주하며, 이를 통해 시퀀스의 섀넌 엔트로피를 추정하는 기반으로 삼는다.

정보이론에서 Fano's inequality 은 오류 확률과 엔트로피 간의 관계를 제시하는 부등식으로, 오류 확률을 분석하는데 활용된다. 이 오류 확률은 최대 확률 (θ) 로 표현이 가능하며, 본 연구에서는 섀넌 엔트로피와 최대 확률의 엔트로피 간의 관계를 분석하는데 사용하고자 한다.

$$h(\theta) + (1-\theta)\log(|\mathcal{X}| - 1) \geq H(X)$$

Fano's inequality 는 분포가 near-uniform 한 경우에 등호가 성립하며, 특히 이진 분포(binary case)에서는 항상 near-uniform 을 만족하기에 등호가 항상 성립한다. 실제 시스템에서 사용되는 난수 소스는 일반적으로 이진 시퀀스로 구성되므로, 해당 부등식을 이용해 최대 확률을 도출한다. 이후, 이 확률값에 음의 로그를 취하여 최소 엔트로피를 추정한다.

* Corresponding author

Ⅲ. 실험 결과

LZ77 압축 알고리즘은 (D, L, C) 형태의 3 중 순서쌍으로 출력되므로, 이를 직접적인 비트 시퀀스로 해석하기에 어려움이 있다. 따라서 압축 길이를 정량화하기 위해서 해당 출력된 순서쌍에 대해 별도의 소스 코딩이 필요하다. 본 연구에서는 이러한 문제를 해결하기 위해 LZ77 기반의 상용 압축 라이브러리인 zlib 을 활용하여 실제 압축 길이를 측정하였다.

실험에 사용된 시퀀스는 두 가지 확률 모델을 따른다. 첫째는 확률 p 로 1 이 발생하는 이진 메모리리스 소스(Binary Memoryless Source; BMS), 둘째는 전이 확률 $p(1|0) = p(0|1) = p$ 을 갖는 이진 마르코프 소스(Binary Markov Chain Source; MCS)이다. 각각의 확률 $p = 0.05$ 부터 0.5 까지 $100{,}000$ 길이의 이진 시퀀스를 생성하였고, 이를 zlib 을 이용해 압축한 후 이를 기반으로 최소 엔트로피를 추정하였다. 또한 기존 국제표준 NIST 에서 제시한 압축 알고리즘을 비교군으로 활용하여 제안 기법의 성능을 평가하였다.

Fig. 1 과 Fig. 2 는 각각 BMS 시퀀스와 MCS 시퀀스에 대한 실험 결과를 나타낸다. 실험 결과에서 알 수 있듯이, 제안한 방법은 실제 이론값에 비해 다소 과추정(overestimate)하는 경향이 있으나, 국제표준에서 제시한 압축 추정기(compression estimator) 비해 오차 편차가 더 작다는 점을 확인할 수 있다.

본 논문에서 제안하는 추정 방법은 LZ77 압축 알고리즘을 기반으로 섀넌 엔트로피를 추정한다. 이 때 도출된 압축률은 이론적인 섀넌 엔트로피보다 큰 값을 가지는 경향이 있다. 이는 LZ77 이 이론적으로 섀넌 엔트로피에 수렴하는 것이 맞지만, 해당 수렴은 데이터 길이와 버퍼 크기가 무한하다는 가정 하에 성립하는 것이며, 실제 유한한 길이(finite length)의 데이터에서는 섀넌 엔트로피보다 다소 큰 값으로 나타나는 것이 일반적이다.

이러한 점이 본 연구에서 제안한 방법이 최소 엔트로피를 과추정하게 되는 주요 원인이며, 사용되는 압축 라이브러리에 따라 데이터의 길이 및 버퍼 크기를 증가시키면 이러한 편차는 줄어들 것으로 기대된다.

Ⅳ. 결론

본 논문에서는 LZ77 압축 알고리즘과 Fano's inequality 를 결합하여 이진 시퀀스의 최소 엔트로피를 추정하는 새로운 방법을 제안하였다. LZ77 의 압축률을 통해 섀넌 엔트로피를 추정하고, 이를 기반으로 최대 발생 확률을 유도함으로써 최소 엔트로피를 추정하였다. 실험 결과, 제안한 방법은 국제 표준인 NIST 에서 제시한 기존 압축 기반 방법에 비해 더 낮은 오차 편차를 갖는 것을 확인하였다.

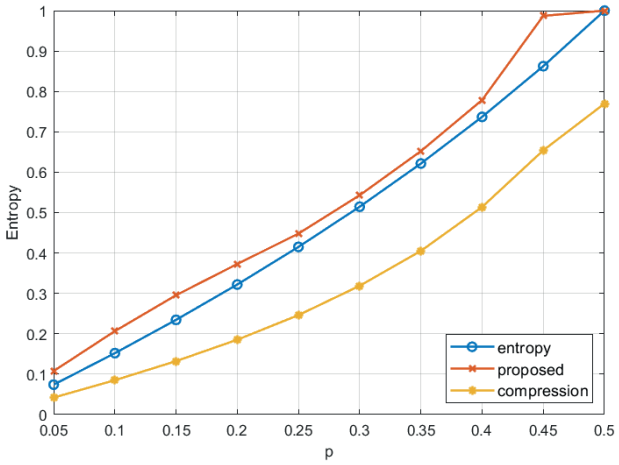

Fig. 1. BMS 에 대한 제안된 추정 방법과 compression 추정 방법 비교.

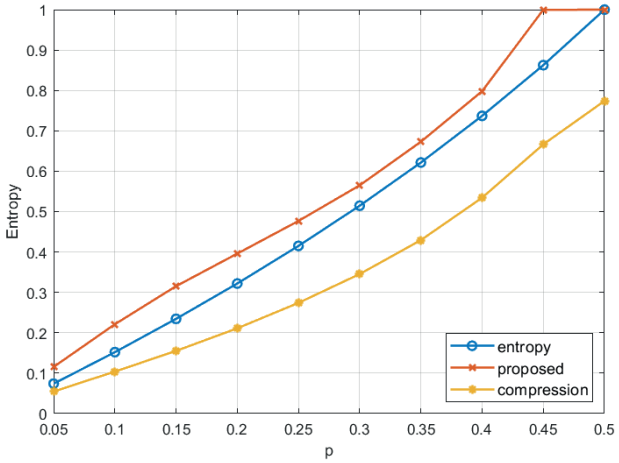

Fig. 2. MCS 에 대한 제안된 추정 방법과 compression 추정 방법 비교.

ACKNOWLEDGMENT

이 논문은 정부(과학기술정보통신부)의 재원으로 정보통신기획평가원의 지원을 받아 수행된 연구임(No. 2024-00399401, 양자안전 보안인프라 전환 및 대양자 복합 안정성 검증기술 개발).

참 고 문 헌

[1] M. S. Turan, E. Barker, J. Kelsey, K. A. McKay, M. L. Baish, and M. Boyle, Recommendation for the entropy sources used for random bit generation, NIST Special Publication 800-90B Std., Jan. 2018.

[2] T. M. Cover and J. A. Thomas, Elements of Information Theory, 2nd ed. Hoboken, NJ, USA: Wiley-Interscience, 2006.