The Journal of The Institute of Internet, Broadcasting and Communication (IIBC)
Vol. 19, No. 5, pp.251-255, Oct. 31, 2019. pISSN 2289-0238, eISSN 2289-0246

https://doi.org/10.7236/JIIBC.2019.19.5.251
JIIBC 2019-5-35

비공개 프로토콜 분류를 위한 특징 추출 알고리즘 비교 연구

A Comparative Study of Feature Extraction Algorithm for unKnown Protocol Classification

정영규*, 정창민**

YoungGiu Jung*, Chang-Min Jeong**

요 약 프로토콜 reverse-engineering 기술은 unknown protocol 의 스펙을 추출하기 위해서 보통 표준화된 방법이 없어서 대부분 수동으로 스펙을 분석하거나 반자동 방식으로 이를 분석한다. 만약 unknown protocol의 근간이 되는 프로토콜을 알 수 있다면, 이를 이용하여 스펙을 분석할 수 있으므로 자동화되고 정확한 분석이 가능할 것이다. 학습되지 않은 프로토콜을 분류하기 위해서는 특징추출은 매우 중요한 단계 중의 하나이다. 본 논문은 기존 프로토콜을 변형한 프로토콜에 대해서 높은 성능을 갖는 분류기를 개발하기 위해서 몇 가지 특징 추출 알고리즘을 제안하고, 프로토콜의 형태 변화에 강인한 특징추출 알고리즘을 제안한다. 성능 검증을 위해서 8개 공개 프로토콜을 대상으로 학습을 수행하고 이를 변형한 프로토콜을 대상으로 성능 측정을 진행하였다.

Abstract On today, Protocol reverse-engineering technique can be used to extract the specification of an unknown protocol. However, there is no standardized method, and in most cases, the extracting process is executed manually or semi-automatically. If the information about the structure of an unknown protocol could be acquired in advance, it would be easy to conduct reverse engineering. the feature extraction is an important step in unknown protocol classification. However, in this paper, we present a comparison several feature extraction techniques and suggests a method of feature extraction algorithm for recognizing unknown protocol. In order to verify the performance of the proposed system, we performed the training using eight open protocols to evaluate the performance using unknown data.

Key Words : Protocol reverse-engineering, feature extraction, Deep Learning, Transformed protocol, moment feature, frequency feature

Ⅰ. 서 론

오늘날 통신 및 사이버 공간에서의 정보 작전은 다양한 형태로 이루어 질 수 있으며 이 중에서 정보공격이 무엇보다도 중요한데 이는 적의 네트워크, 정보 시스템, 무기체계 등의 취약점을 분석하고 침투를 통하여 정보 흐름을 탐지, 추출하고 이를 위조, 변조하여 적의 원활한 정보흐름을 차단하거나 마비 파괴하여 적 전력을 약화시키

*정회원, 인하대학교 컴퓨터공학과
**정회원, 국방과학연구소
접수일자 2019년 8월 7일, 수정완료 2019년 9월 7일
게재확정일자 2019년 10월 4일

Received: 7 August, 2019 / Revised: 7 September, 2019 / Accepted: 4 October, 2019
**Corresponding Author: rerajung@gmail.com
Dept. of Computer Engineering, Inha University, Korea

는 기술이다. 이러한 정보공격을 위해서는 적의 통신 패킷으로부터 프로토콜을 분석할 수 있으면, 다양한 형태의 정보공격이 가능하게 된다[1]. 본 논문은 군의 사이버 공격/방어 무기 개발을 위해 수집된 적 통신 패킷을 분석하여 프로토콜의 종류를 분류하는 기술개발에 가장 중요한 기술 중의 하나인 특징추출 알고리즘에 대해서 분석하고자 한다.

본 논문은 공개 프로토콜로 학습 후 변형 프로토콜이 공개 프로토콜과 얼마나 유사한지를 추정하는 기술로써 기존 프로토콜 분류기와는 다른 영역의 연구분야이다. 본 연구에서 가장 중요한 것중의 하나가 특징추출 알고리즘이다. 프로토콜 분류를 위한 특징추출 알고리즘 연구는 프로토콜 identification 분야에서 많은 연구가 이루어졌다.

J. Zhang[2]은 Packet의 수, Packet Size, Inter-Packet Time 등을 특징으로 사용하였으며, Rongqiang Lin[3]은 client-to-server 간의 패킷수, 최대 패킷 크기, 최소 패킷크기, client-to-server 패킷간의 표준편차 등을 특징으로 사용했다. H.L Yu[4]은 source/destination IP, source/destination port 등을 특징으로 사용하였으며, A McGregor[5]은 packet 길이 통계, inter-arrival 통계, byte 수, connection 시간 등을 특징으로 사용하여 프로토콜을 분류하였다. 본 논문은 패킷의 흐름을 제외하고 패킷 자체에 포함된 정보만을 분석하는 특징추출 알고리즘을 개발하고 이를 근간으로 비공개 프로토콜 분류 시스템을 개발하고자 한다.

II. 특징추출 알고리즘

1. 통계기반 특징추출 알고리즘

통계기반 특징은 일정한 수의 패킷을 수집한 후 패킷 내 필드 값의 통계를 계산하여 이를 특징으로 사용하는 방법이다. 통계기반 특징추출은 n개의 패킷에 대해서 모멘트와 빈도를 특징으로 추출한다. 그림 1은 통계기반 특징추출 알고리즘의 순서도이다.

모멘트는[6] 각 필드의 평균, 분산, 왜도, 첨도를 추출하여 특징으로 생성하였으며, 빈도는 각 필드에서 빈도를 생성한 후 최고 빈도 값을 특징으로 사용한다. 그림 2는 두 특징추출 알고리즘의 개념도이다.

그림 3은 개별 프로토콜의 모멘트 및 빈도 특징 추출 결과이다. 그림에서 왼쪽은 빈도 특징이고 오른쪽은 모멘트에서 평균 특징추출 결과이다.

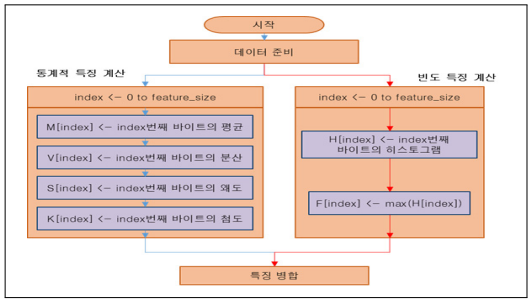

그림 1. 통계기반 특징추출 알고리즘 순서도
Fig. 1. Flowchart of Statistical feature Extraction Algorithm

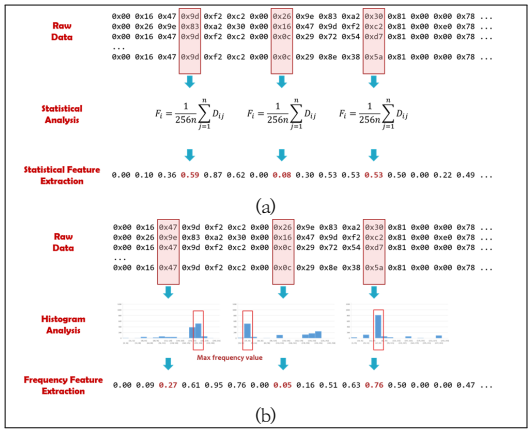

그림 2. (a) 모멘트 특징 추출, (b) 빈도 특징추출
Fig. 2. (a) Moment Feature extraction, (b) Frequency feature extraction

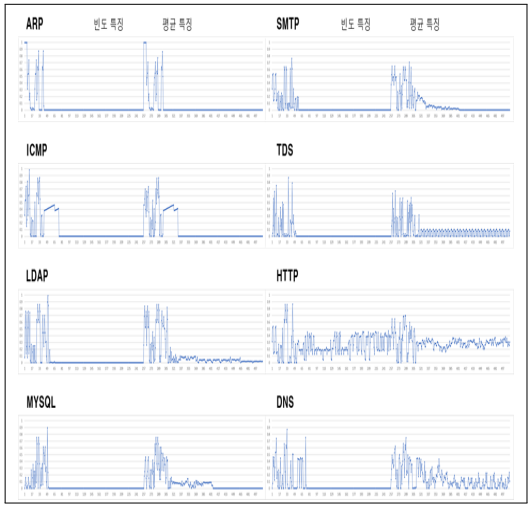

그림 3. 모멘트 및 빈도 특징추출 결과 비교
Fig. 3. Compare of Moment and frequency feature

2. AutoEncoder 기반 특징추출 알고리즘

AutoEncoder 는 입력 값을 출력값으로 복사하는 방법으로 특징을 생성하는 비지도 학습 신경망 모델이다. 일반적으로 입력 개수와 출력 개수가 동일한데 이때 hidden 계층 값이 특징으로 사용된다. 그림 4는 패킷의 원본 데이터로부터 학습 모델을 이용하여 특징을 생성하기 위한 AutoEncoder의 구조이다.

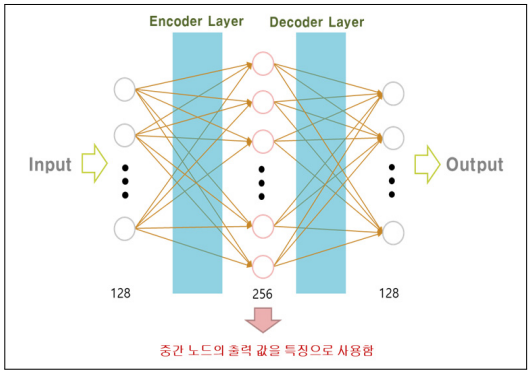

그림 4. AutoEncoder 특징추출의 구조
Fig. 4. Structure of AutoEncoder

본 논문에서 제안된 AutoEncoder의 구조는 한 개의 히든레이어로 구성되어진다. 제안된 AutoEncoder의 구조는 입력 데이터는 128이고 이를 256개의 특징 공간으로 확장하는 구조를 가진다. 그리고 본 논문에서 적용한 AutoEocoder의 학습 방법은 아래 수식과 같다. L은 손실 함수인데 일반적으로 평균제곱오차를 사용한다.

$$\hat{\theta} = \arg min \sum_{i}^{n} L(x_i, g(f(x_i))) \quad (1)$$

$$L(x_i, g(f(x_i))) = \| x_i - g(f(x_i)) \|_2^2 \quad (2)$$

3. PCA 기반 특징추출 알고리즘

통계기반과 AutoEncoder 기반 특징추출 알고리즘을 적용할 때 가장 큰 문제 중의 하나는 패킷 내 데이터의 순서가 바뀐 변형 프로토콜를 테스트 데이터로 사용할 때 성능 저하가 발생하게 된다는 것이다.

본 논문에서 테스트 데이터로 사용하는 unknown 프로토콜 (Transformed 프로토콜)은 기존 프로토콜의 정보들 중 에서 일부 값의 위치를 바꾸거나, 일부를 삭제하거나, 일부 데이터를 삽입한 프로토콜로 정의된다. 그림 5는 변형 프로토콜의 한 형태이다. 그림 5는 변형 프로토콜의 예이다. 그림5에서 (a)는 원본 ICMP 이고 (b)는 패킷에서 일부 데이터의 위치를 변경한 프로토콜이다.

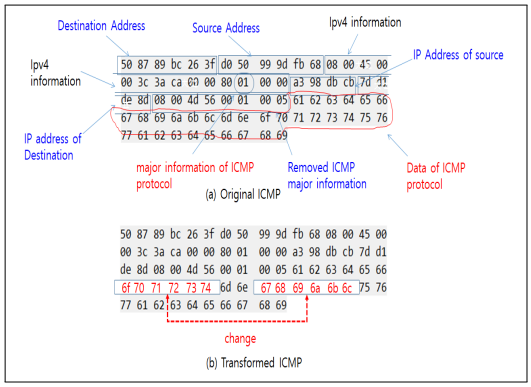

그림 5. (a) Original Protocol, (b) Transformed Protocol 의 예
Fig. 5. Example of (a) Original Protocol, (b) Transformed Protocol

본 논문은 이러한 문제를 해결하기 위해서 프로토콜내 데이터의 위치 변화에 강인한 특징 생성을 위해서 Principal Component Analysis (PCA)을 이용하여 순서에 상관없는 특징을 생성하는 방법을 제안한다. 그림 6은 기존 특징과 PCA를 결합한 특징추출 알고리즘 구성도이다.

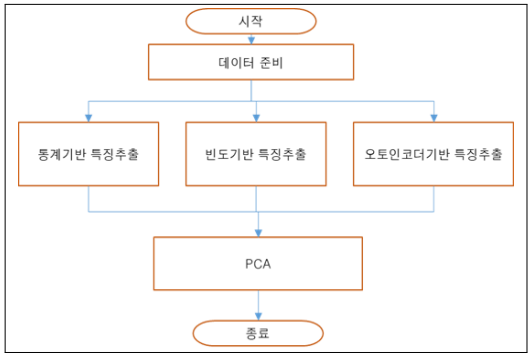

그림 6. PCA기반 특징추출 알고리즘 구성도
Fig. 6. PCA based on feature extraction algorithm

Ⅳ. 실험 및 결과

제안된 특징추출 알고리즘의 성능 검증을 위해서 딥러닝 알고리즘 중의 하나인 Deep Belief Networks[7]를

연동하여 인식 성능을 측정 한다[8][9]. 실험 대상 프로토콜은 ARP, ICMP, LDAP, MYSQL, TDS, HTTP, DNS로 8종의 프로토콜을 대상으로 수행한다. 그림 7은 특징별 기존 프로토콜에 대한 성능측정 결과이다. 통계와 빈도 특징을 함께 사용한 경우에 높은 성능을 보임을 알 수 있다.

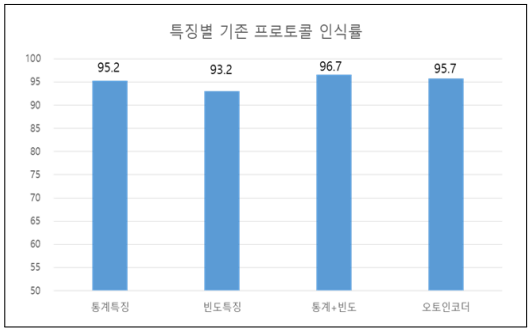

그림 7. 특징 추출 알고리즘 별 성능 비교 결과
Fig. 7. Performance comparison by each feature extraction algorithm

다음은 변형 프로토콜에 따른 성능 검증을 수행하였다. 변형 프로토콜은 크게 20%, 40%, 60%, 80%, 100%의 변형률을 적용하여 테스트데이터를 생성하였다. 변형 프로토콜은 PCA를 적용한 경우와 적용하지 않은 경우로 나누어 분석을 했으며, PCA를 적용한 경우 변형률이 낮을때는 기존 보다 높은 성능을 보였으며, 변형률이 높아질때에는 통계적인 기법보다 낮은 성능으로 나타났다.

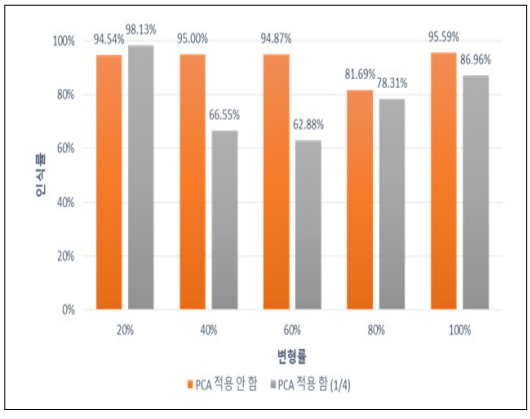

그림 8. 변형률에 따른 비공개 프로토콜 성능 비교
Fig. 8. Performance comparison by transformed protocol

V. 결 론

본 논문은 비공개 프로토콜 분류를 위한 특징추출 알고리즘을 개발하고 제안된 특징추출 알고리즘의 장 단점을 기술한다. 비공개 프로토콜 분류를 위한 특징추출 알고리즘으로 통계기반 특징추출 알고리즘과 빈도기반 특징추출 알고리즘 그리고 원본 데이터로부터 특징을 생성하는 오토인코더를 이용한 특징추출 알고리즘의 성능을 비교하였다. 제안된 특징추출 알고리즘의 성능 검증을 위해서 8개의 프로토콜을 대상으로 성능 검증을 수행하였으며, 성능 검증 결과 통계 특징과 빈도 특징을 함께 사용하는 것이 가장 높은 성능을 보임을 알 수 있었다.

이러한 연구를 기반으로 비공개 프로토콜에 대해서 성능 검증을 수행하였다. 비공개 프로토콜은 8개 프로토콜을 변형한 프로토콜로써, 학습은 공개 프로토콜로 수행하고 테스트는 비공개 프로토콜을 이용하여 성능을 비교하였다. 통계 와 빈도 특징의 경우 약간의 프로토콜 변형에도 성능 저하가 발생하였으며, 이를 극복하기 위해서 PCA기반 특징추출 알고리즘을 개발하였다. 제안된 알고리즘의 비공개 프로토콜에 대한 실험 결과 PCA를 적용한 경우 변형률이 낮을 때에는 기존 방법에 비해서 높은 성능이 나옴을 알 수 있었다.

References

[1] J. Zhang, X. Chen, Y. Xiang, W. Zhou, J. Wu, “Trend and Prospect of Security System Technology for Netwrok,” The Journal of The Institute of Internet, Broadcasting and Communication, vol. 18, no. 5, ,pp.1-8, Oct. 31, 2018.
dio:10.7236/JIIBC.2018.18.5.1

[2] J. Zhang, X. Chen, Y. Xiang, W. Zhou, J. Wu, “Robust Network Traffic Classification,” IEEE/ACM TRANSACTIONS ON NETWORKING, VOL. 23, NO. 4,pp.1257- 1270, 2015.

[3] Rongqiang Lin, Ou Li, Qing Li, and Yan Liu “Unknown Network Protocol Classification Method based on Semi-Supervised Learning,” 2015 IEEE International Conference on Computer and Communications (ICCC) ,pp.300- 308, 2015

[4] H. L Yu, Y. Zhao, G. Xiong, L. Guo, Z. Li, Y. Wang, “POSTER: Mining Elephant Applications in Unknown Traffic by ServiceClustering,” Proceedings of the 2014 ACM SIGSAC Conference on Computer and Communications Security. ACM, pp. 1532-1534, 2014.

[5] A. McGregor, M. Hall, P. Lorier, and J. Brunskill, “Flow clustering using machine learning techniques,” in

Proc. Passive and Active Measurement Workshop (PAM2004), Antibes Juan-les-Pins, France, April 2004.

[6] Spanos, Aris. Probability Theory and Statistical Inference. New York: Cambridge University Press. pp. 109-130. ISBN 0-521-42408-9, 1999

[7] G. Hinton, S. Osindero, and Y. Teh, "A fast learning algorithm for deep belief nets," Neural Comput., vol. 18, no. 7, pp. 1527-1554, 2006.

[8] Park, Jong-Jun, and Chun-Ki Kwon. "A Consistency Study of CNN's Learning to Recognize Korean Finger Number using sEMG Signals," Journal of the Korea Academia-Industrial cooperation Society, vol. 19, no. 10, Oct. 2018, pp. 523-529, doi:10.5762/KAIS.2018.19.10.523

[9] Kwihoon Kim and Bangwon Seo, "Intelligent Construction Video Management System Based on Edge Computing Using Deep Learning," The Journal of Korean Institute of Information Technology, vol. 17, no. 7, Sep. 2019, pp. 55-63, doi:10.14801/jkiit.2019.17.7.55

저 자 소 개

Young Giu Jung(정회원)

- 2008년 경북대학교 컴퓨터공학과 공학박사
- 2011년 ~ 현재 YM-나을텍 대표
- 2018년 ~ 현재 인하대학교 컴퓨터공학과 겸임교수
- 관심분야 : 컴퓨터비젼, 음성인식, 표적추적, 프로토콜 인식

Chang-Min Jeong(정회원)

- 2018년 경북대학교 컴퓨터공학과 공학박사
- 2015년 ~ 현재 국방과학연구소 책임연구원
- 관심분야 : 사이버전, 프로토콜 식별, 레이더인식