개선된 데이터 마이닝 기법을 이용한 악성코드 탐지 연구 동향

봉기정, 김종현*
한국전자통신연구원, *한국전자통신연구원
bkj8797@etri.re.kr, *jhk@etri.re.kr

A Survey on Malware Detection using Enhanced Data Mining Techniques

Bong Ki Jung, Kim Jong Hyun*
ETRI., *ETRI.

요 약

최근 악성코드를 이용한 해킹 기술이 지능화되고 고도화됨에 따라, 여러 조직과 기관의 악성코드 방어 체계가 한계에 다다르고 있다. 기존의 패턴 기반 악성코드 탐지 방식은 잘 알려진 악성코드에 대해서는 뛰어난 탐지율을 보이지만, 알려지지 않았거나 새로운 패턴의 악성코드에 대해서는 좋은 성능을 내지 못한다는 단점이 있다. 이에 대한 방안으로, 데이터 마이닝 기반의 탐지 방식은 알려지지 않은 악성코드에 대해 패턴 기반의 탐지 방식보다 뛰어난 성능을 보여준다. 본 논문에서는 다양한 데이터 마이닝 기법들과 해당 기법들을 개선한 선행 연구들을 소개한다. 또한, 선행 연구들에 존재하는 논점과 향후 연구의 요구사항들을 다룬다.

Ⅰ. 서 론

악성코드 탐지, 분석, 방어 기술은 보안 업계가 풀어야 할 영원한 숙제이다. 패킹, 난독화 등의 기술을 이용한 악성코드 제작 기술이 심화 되고 있고, 악성코드를 이용한 해킹 기술 또한 지능화되고 고도화되고 있다. 이에 따라 악성코드 방어 체계를 구축하기 위한 연구가 활발히 진행되고 있다.

기존의 signature 기반 악성코드 탐지 방식은 변형되었거나 기존에 알려진 방식이 아닌(이하 unknown이라 칭함.) 악성코드를 탐지하기 힘든 어려움이 있다. 이를 보완하기 위해, 데이터 마이닝 기반 악성코드 탐지의 많은 선행 연구가 진행되었다.

데이터 마이닝을 이용한 악성코드 탐지 방식은 unknown 악성코드에 대하여 탐지율이 좋은 장점이 있다. 데이터 마이닝을 이용한 방식에는 다양한 기법이 있으며, 주로 Naive Bayes(NB), Decision Tree(DT), Random Forest(RF), Support Vector Machine(SVM), k-Nearest Neighbor(kNN) 알고리즘 등이 높은 성능을 내는 것으로 알려져 있다. 본 논문에서는 이러한 기존 데이터 마이닝 기반의 탐지 방식을 개선하여 더 효율적으로 악성코드를 탐지하는 방안을 제시한 연구사례를 조사하였다.

Ⅱ. 관련 연구

기존의 데이터 마이닝 기법을 이용한 악성코드 탐지 기법을 다양한 방식으로 개선한 연구들은 다음과 같다.

관련 연구 [1]에서는 Raw Data에 TF-IDF 기법을 이용해 Feature를 추출하고, 해당 데이터에 MIFS를 이용해 가장 영향력 있는 Feature를 선정하였다. 이후 해당 데이터에 DT, MLP, SVM(OCSVM), PNN, GMDH 기법을 적용하여 accuracy를 비교하였다. 해당 실험에서는 대부분 70%의 accuracy를 보였으며, OCSVM의 경우 약 98.5%~100%의 accuracy를 보이기도 했다.

악성코드의 대부분을 차지하는 소규모 실행 파일의 신종 및 변종을 식별해낼 수 없었던 문제에 대해, 관련 연구[2]에서는 소규모 악성코드를 온라인 분석, 정적 분석, 동적 분석을 혼합하여 메타 데이터를 생성한 뒤 이들에 NB, SVM, DT, RF, ANN 기법을 적용하여 accuracy를 비교하였다. 해당 실험에서는 RF가 약 89%의 accuracy를 보였다.

관련 연구 [3]에서는 악성코드의 동작을 tracing 하는 프로그램을 이용해 악성코드가 숨기려고 하는 정보를 후킹 하여, API와 같이 자신의 코드나 동작을 숨기는 악성코드에 대해서도 데이터 마이닝 기법을 적용하여 분석했다. 해당 실험에서는 NB, DT(J48), SVM 기법을 사용하여 accuracy를 비교하였으며, J48과 NB가 약 95%의 accuracy를 보였다.

관련 연구 [4]에서는 Feature Selection 과정에 Opcode 4-gram을 이용하였다. 이후 해당 데이터에 SVM, NB, DT, J48, kNN, RF 기법을 적용하여 accuracy를 비교하였다. 해당 실험에서는 SVM이 약 98.2%의 accuracy를 보였다.

관련 연구 [5]에서는 실행 파일의 string 문자열을 추출해 TF-IDF 기법을 적용한 뒤, 해당 데이터를 벡터화하여 Feature로 활용하였다. 벡터 간의 Cosine Similarity를 이용하여 kNN 알고리즘을 적용하거나, DNN 알고리즘을 적용하여 악성코드 탐지 및 분류 실험을 진행하였다. 해당 실험에서는 윈도우 기준 약 90%, 리눅스 기준 약 97%의 accuracy를 보였다.

관련 연구들의 Feature Selection 기법과 accuracy를 정리하여 [표 1]에 나타내었다. 관련 연구 [1]에서 가장 높은 accuracy를 보였고, SVM 알고리즘이 대체로 높은 효율을 보였다.

Ⅲ. 도전 과제

이번 장에서는 선행되었던 연구들의 향후 개선점과 이를 토대로 한 앞으로 진행될 연구에서의 요구사항을 다룬다. 위에서 다루었던 연구들은 모두 훌륭한 실험 결과를 보여주었으나, 향후 개선하거나 새로 다룰 수 있는 부분은 분명 존재한다.

첫 번째로, 텍스트 마이닝 기법 사용 시, 다른 텍스트 마이닝 기법의

Reference Number	Feature Selection Technique	Accuracy
[1]	TF-IDF, MIFS	**98.5%(OCSVM)**, 70.27%(SVM), 66.13%(PNN), 73.43%(DT), 70.47%(GMDH), 70.05%(MLP)
[2]	Binary Features of Static/Dynamic Analytics	73%(NB), 87%(SVM), 82%(DT), **88%(RF)**, 84%(ANN)
[3]	API frequency	**95%(J48)**, 95%(NB), 89%(SVM)
[4]	Opcode 4-gram	**98.2%(SVM)**, 94.5%(NB), 94.5%(kNN), 89.0%(DT), 93.3%(J48), 96.9%(RF)
[5]	TF-IDF	90.99%(DNN, Windows), 89.88%(kNN, Windows), **97.45%(DNN, Linux)**, 96.70%(kNN, Linux)

[표 1] 관련 연구 별 사용 기법과 Accuracy

채택도 고려해볼 수 있다. 관련 연구 [1][5]에서 사용한 TF-IDF는 유니크한 단어를 추출하는데 좋은 기법이지만, 악성코드에서는 특정 단어보다 특정 문맥이 더 중요한 경우가 많다. 이 경우 TF-IDF 값을 추출하여 활용했을 때 좋은 성능을 도출하기 힘들다는 한계점이 있기 때문에, BERT 등 다른 언어 모델을 사용하는 측면을 고려해볼 수 있다.

두 번째로, 데이터의 중복을 확인하는 방법이 있다. 관련 연구 [1][4][5]에서 보여준 것처럼 accuracy가 100%에 가까운 모델은 분명 훌륭한 모델링을 통해 굉장히 좋은 성능을 보여주고 있지만, 데이터의 중복으로 인해 지나치게 높은 accuracy가 도출되었을 가능성이 있다.

세 번째로, 악성코드 탐지 모델 학습 시 많은 양의 학습 데이터를 사용할 수 있다. 악성코드는 웜, 바이러스, 트로이목마 등 다양한 패밀리가 존재하기 때문에, 하나의 모델에 적용하기 위해서는 각 패밀리 별로 많은 양의 데이터를 학습시켜야 정확한 효율을 추출할 수 있으며, 높은 정확도를 확보할 수 있다. 관련 연구 [5]와 같이 악성코드 그룹 분류에 비교적 적은 양의 데이터를 사용하는 경우, k-fold 교차 검증 방식이나 GAN(Generative Adversarial Network) 모델 등을 사용하여 보완하는 방법이 있다.

네 번째로, 가상 환경 혹은 모니터링 툴을 인식하고 동작하지 않는 Anti-VM 등의 기법이 적용된 악성코드에 대한 대책이 필요하다. 관련 연구 [3]과 같이 가상 환경에서 모니터링 툴을 이용해 동적 분석 하는 경우, 해당 악성코드가 행하는 코드 인젝션이나 프로세스 생성 등의 동작을 탐지하지 못하는 경우가 발생할 수 있다. 따라서 Anti-VM 기법을 방지하는 방안을 적용하여 동적 분석을 진행하는 방향을 고려할 수 있다.

위에서 다룬 선행 연구의 개선점이나 기존에 존재하던 논점들로부터 향후 진행될 연구의 요구사항을 도출해 정리하였다. 먼저, 사용할 수 있는 양질의 학습 데이터가 많아져야 한다. 현재 국내에서 주로 사용되는 악성코드 데이터는 대부분 공통적이며 그 수가 한정되어 있다. 백신 기관이나 그와 관련된 기관에서는 연구에 사용할 수 있는 악성코드 데이터가 많지만, 일반 학생들이나 타 기관에서는 양질의 데이터 확보에 어려움이 있다. 머신러닝 혹은 딥러닝 기반의 연구에 있어 양질의 데이터는 굉장히 중요한 요소인데, 이러한 데이터를 활용한다면 연구의 질과 속도가 향상될 것이다. 추가로, 사용한 데이터의 검증 또한 필수적이다. 일부 연구에서 지나치게 높은 accuracy가 측정되는 경우가 있는데, 이 경우 test dataset과 train dataset의 중복 검증을 수행할 필요가 있다. 또한, 악성코드 분석/탐지 분야의 지속적인 개선이 필요하다. 패킹, 난독화, Anti-VM 등 악성코드 분석을 어렵게 하는 기술들이 발전하고 있고, 새로운 형태의 악성코드가 등장하고 있다. 이에 따라 이러한 악성코드의 탐지에 사용되는 모델의 적용에도 어려움이 따른다. 마지막으로, 악성코드의 Labeling에 관한 문제를 해결해야 한다. 최근의 악성코드는 여러 가지 기능을 포함하기도 해 하나의 Label로 표현하는 데에 어려움이 있다. 또한, 백신사마다 Labeling을 다르게 하고 있어, 악성코드 그룹 분류 등에 큰 걸림돌이 된다.

Ⅳ. 결론

난독화, 패킹 등 악성코드 제작 기술과 stealth 등 악성코드의 활용 기술이 고도화되고 있다. 이에 기존의 악성코드 탐지 방식으로는 탐지하지 못하는 악성코드로 인해 많은 기업과 기관들이 큰 피해를 받고 있다. 악성코드로부터 의미 있는 feature를 추출하여 이를 데이터 마이닝 기법에 적용해 악성코드를 탐지하는 데이터 마이닝 기반 탐지기법이 존재하며, 이러한 기법들을 개선하려는 많은 연구가 진행되었다. 2장에서는 데이터 마이닝을 이용해 악성코드를 탐지하는 기법을 다양한 방식으로 발전시킨 논문들을 살펴보았다. 대부분 좋은 성능을 낸다고 알려진 SVM, DT, RF, NB 등의 알고리즘을 사용했으며, 일부 연구에서는 TF-IDF, GMDH 등의 기법을 활용하기도 했다. 또한, 3장에서는 개선된 연구에서 향후 더욱 개선할 수 있는 사항은 무엇인지, 앞으로의 연구 요구사항은 무엇인지를 다루었다. 향후 연구에서 이러한 요구사항들을 해결한다면, 악성코드 분야 연구에 큰 발전이 있을 것으로 기대된다.

ACKNOWLEDGMENT

이 논문은 2021년도 정부(과학기술정보통신부)의 재원으로 정보통신기획평가원의 지원을 받아 수행된 연구임 (No.2019-0-00026, 지능화된 악성코드 위협으로부터 ICT 인프라 보호)

참 고 문 헌

[1] Sundarkumar, G. G., and Ravi, V. "Malware Detection by Text and Data Mining," Computational Intelligence and Computing Research (ICCIC), IEEE International Conference on, pp. 1-6, Dec. 2013.

[2] 이택현, 국광호 "데이터 마이닝 기법을 이용한 소규모 악성코드 탐지에 관한 연구," 융합보안논문지, 제19권 제1호, pp. 11-1, 2019.

[3] Fan, C. I., Hsiao, H. W., Chou, C. H., and Tseng, Y. F. "Malware Detection Systems Based on API Log Data Mining," IEEE 39th Annu. Comput. Softw. Appl. Conf., pp. 255 - 260, Jul. 2015.

[4] Samantray, O. P., Tripathy, S. N., Das, S. K. "A Data Mining Based Malware Detection Model using Distinct API Call Sequences," International Journal of Innovative Technology and Exploring Engineering (IJITEE), Volume-8 Issue-7, May. 2019.

[5] 하지희, 이태진. "문자열 정보를 활용한 텍스트 마이닝 기반 악성코드 분석 기술 연구," 인터넷정보학회논문지 21.1 45-55, 2020.

[6] Bhaya, W., and Ali, M. "Review on Malware and Malware Detection Using Data Mining Techniques," Journal of University of Babylon, 25(5), 1585 - 1601, 2017.

[7] Souri, A., and R, Hosseini. "A state-of-the-art survey of malware detection approaches using data mining techniques," Human-centric Computing and Information Sciences 8, no. 1, 2018.

[8] Poudyal, S., Akhtar, Z., Dasgupta, D., and Gupta, K. D. "Malware Analytics: Review of Data Mining, Machine Learning and Big Data Perspectives," IEEE Symposium Series on Computational Intelligence (SSCI), pages 649 - 656, 2019.