ScienceDirect

Computers & Security

Volume 81, March 2019, Pages 123-147

# Survey of machine learning techniques for malware analysis

Daniele Ucci [a], Leonardo Aniello [b], Roberto Baldoni [a]

Show more



https://doi.org/10.1016/j.cose.2018.11.001

Get rights and content

## Abstract

Coping with malware is getting more and more challenging, given their relentless growth in complexity and volume. One of the most common approaches in literature is using machine learning techniques, to automatically learn models and patterns behind such complexity, and to develop technologies to keep pace with malware evolution. This survey aims at providing an overview on the way machine learning has been used so far in the context of malware analysis in Windows environments, i.e. for the analysis of Portable Executables. We systematize surveyed papers according to their objectives (i.e., the *expected output*), what information about malware they specifically use (i.e., the *features*), and what machine learning techniques they employ (i.e., what *algorithm* is used to process the input and produce the output). We also outline a number of issues and challenges, including those concerning the used *datasets*, and identify the main current topical trends and how to possibly advance them. In particular, we introduce the novel concept of *malware analysis economics*, regarding the study of existing trade-offs among key metrics, such as analysis accuracy and economical costs.

## Introduction

Despite the significant improvement of cyber security mechanisms and their continuous evolution, malware are still among the most effective threats in the cyber space. Malware analysis applies techniques from several different fields, such as program analysis and network analysis, for the study of malicious samples to develop a deeper understanding on several aspects, including their behaviour and how they evolve over time. Within the unceasing arms race between malware developers and analysts, each advance in security technology is usually promptly followed by a corresponding evasion. Part of the effectiveness of novel defensive measures depends on what properties they leverage on. For example, a detection rule based on the MD5 hash of a known malware can be easily eluded by applying standard techniques like obfuscation, or more advanced approaches such as *polymorphism* or *metamorphism*. For a comprehensive

review of these techniques, refer to Ye et al. (2017). These methods change the binary of the malware, and thus its hash, but leave its behaviour unmodified. On the other side, developing detection rules that capture the semantics of a malicious sample is much more difficult to circumvent, because malware developers should apply more complex modifications. A major goal of malware analysis is to capture additional properties to be used to improve security measures and make evasion as hard as possible. Machine learning is a natural choice to support such a process of knowledge extraction. Indeed, many works in literature have taken this direction, with a variety of approaches, objectives and results.

This survey aims at reviewing and systematising existing literature where machine learning is used to support malware analysis of Windows executables, i.e. Portable Executables (PEs). The intended audience of this survey includes any security analysts, i.e. security-minded reverse engineer or software developer, who may benefit from applying machine learning to automate part of malware analysis operations and make the workload more tractable. Although mobile malware represents an ever growing threat, Windows largely remains the preferred target(AV-TEST, 2017) among all the existing platforms. Malware analysis techniques for PEs are slightly different from those for Android apps because there are significant dissimilarities on how operating system and applications work. As a matter of fact, literature papers on malware analysis commonly point out what specific platform they target, so we specifically focus on works that consider the analysis of PEs. 64 recent papers have been selected on the basis of their bibliographic significance, reviewed and systematised according to a taxonomy with three fundamental dimensions: (i) the specific *objective* of the analysis, (ii) what types of *features* extracted from PEs they consider and (iii) what machine learning *algorithms* they use. We distinguish three main objectives: *malware detection, malware similarity analysis* and *malware category detection*. PE features have been grouped in eight types: *byte sequences, APIs/system calls, opcodes, network, file system, CPU registers, PE file characteristics* and *strings*. Machine learning algorithms have been categorized depending on whether the learning is *supervised, unsupervised* or *semi-supervised*. The characterisation of surveyed papers according to such taxonomy allows to spot research directions that have not been investigated yet, such as the impact of particular combination of features on analysis accuracy. The analysis of such a large literature leads to single out three main issues to address. The first concerns overcoming modern anti-analysis techniques such as encryption. The second regards the inaccuracy of malware behaviour modelling due to the choice of what operations of the sample are considered for the analysis. The third is about the obsolescence and unavailability of the datasets used in the evaluation, which affect the significance of obtained results and their reproducibility. In this respect, we propose a few guidelines to prepare suitable benchmarks for malware analysis through machine learning. We also identify a number of topical trends that we consider worth to be investigated more in detail, such as malware attribution and triage. Furthermore, we introduce the novel concept of *malware analysis economics*, regarding the existing trade-offs between analysis accuracy, time and cost, which should be taken into account when designing a malware analysis environment.

The novel contributions of this work are

- the definition of a taxonomy to synthesise the state of the art on machine learning for malware analysis of PEs;
- a detailed comparative analysis of existing literature on that topic, structured according to the proposed taxonomy, which highlights possible new research directions;
- the determination of present main issues and challenges on that subject, and the proposal of high-level directions to investigate to overcome them;
- the identification of a number of topical trends on machine learning for malware analysis of PEs, with general guidelines on how to advance them;

- the definition of the novel concept of malware analysis economics.

The rest of the paper is structured as follows. Related work are described in Section2. Section3 presents the taxonomy we propose to organise reviewed malware analysis approaches based on machine learning, which are then characterised according to such a taxonomy in Section4. From this characterisation, current issues and challenges are pointed out in Section5. Section6 highlights topical trends and how to advance them. Malware analysis economics is introduced in Section7. Finally, conclusions and future works are presented in Section8.

## Section snippets

### Related work

Other academic works have already addressed the problem of surveying contributions on the usage of machine learning techniques for malware analysis. The survey written by Shabtai et al. (2009) is the first one on this topic. It specifically deals with how classifiers are used on static features to detect malware. As most of the other surveys mentioned in this section, the main difference with our work is that our scope is wider as we target other objectives besides malware detection, such as…

### Taxonomy of machine learning techniques for malware analysis

This section introduces the taxonomy on how machine learning is used for malware analysis in the reviewed papers. We identify three major dimensions along which surveyed works can be conveniently organised. The first one characterises the final *objective* of the analysis, e.g. malware detection. The second dimension describes the *features* that the analysis is based on in terms of how they are extracted, e.g. through dynamic analysis, and what features are considered, e.g. CPU registers. Finally, …

### Characterization of surveyed papers

In this section we characterize each reviewed paper on the basis of analysis objective, used machine learning algorithm and features. Several details are also reported on the dataset used for the evaluation, including whether it is publicly available (*Public* column), where samples have been collected from (*Source* column) and whether the specific set of samples considered for the experiment is available (*Available* column). Indeed, many works declare they do not use all the executables in the…

### Issues and challenges

Based on the characterization detailed in Section4, this section identifies the main issues and challenges of surveyed papers. In the specific, the main problems regard the usage of anti-analysis techniques by malware (Section5.1), what operation set to consider(Section5.2) and used dataset(Section5.3).…

### Topical trends

This section outlines a list of topical trends in malware analysis, i.e. topics that are currently being investigated but have not reached the same level of maturity of the other areas described in previous sections.…

## Malware analysis economics

Analysing samples through machine learning techniques requires complex computations for extracting desired features and running chosen algorithms. The time complexity of these computations has to be carefully taken into account to ensure they complete fast enough to keep pace with the speed new malware are developed. Space complexity has to be considered as well, indeed feature space can easily become excessively large (e.g., using n-grams), and also the memory required by machine learning…

## Conclusion

We presented a survey on existing literature on malware analysis through machine learning techniques. There are five main contributions of our work. First, we proposed an organization of reviewed works according to three orthogonal dimensions: *the objective of the analysis, the type of features extracted from samples*, the *machine learning algorithms used to process these features*. Such characterization provides an overview on how machine learning algorithms can be employed in malware analysis,…


## Acknowledgment

This work has been partially supported by a grant of the Italian Presidency of Ministry Council and by the Laboratorio Nazionale of Cyber Security of the CINI (Consorzio Interuniversitario Nazionale Informatica).…



**Daniele Ucci** is a Ph.D. student in Engineering in Computer Science at Department of Computer, Control, and Management Engineering “Antonio Ruberti” at Sapienza University of Rome. He received the master degree with honors in Computer Engineering in Computer Science in 2014 A.Y.. His research interests mainly focus on Big Data and information security and privacy, with special regard to malware analysis. During his master thesis, he has investigated topics related to business intelligence and…



Recommended articles


## References (107)


P.M. Comar *et al.*
Combining supervised and unsupervised learning for zero-day malware detection
Proceedings of the 32nd annual IEEE international conference on computer communications (INFOCOM) (2013)

M. Graziano *et al.*
Needles in a haystack: Mining information from public dynamic analysis sandboxes for malware intelligence
Proceedings of the 24th USENIX Security Symposium(2015)

JangJ. *et al.*
Bitshred: feature hashing malware for scalable triage and semantic analysis
Computer and communications security(2011)

Offensive Computing. http://www.offensivecomputing.net. Accessed:...

M. Polino *et al.*
Jackdaw: towards automatic reverse engineering of large datasets of binaries
Detection of intrusions and malware, and vulnerability assessment(2015)

I. Santos *et al.*
International symposium on distributed computing and artificial intelligence
(2011)

A. Souri *et al.*
A state-of-the-art survey of malware detection approaches using data mining techniques
Hum Cent Comput Inf Sci (2018)

M. Ahmadi *et al.*
Novel feature extraction, selection and fusion for effective malware family classification
CoRR (2015)

F. Ahmed *et al.*
Using spatio-temporal information in APi calls with machine learning algorithms for malware detection
Proceedings of the 2nd ACM workshop on security and artificial intelligence(2009)

F.E. Allen
Control flow analysis
Proceedings of a symposium on compiler optimization(1970)

View more references

## Cited by (364)

Android malware detection and identification frameworks by leveraging the machine and deep learning techniques: A comprehensive review
2024, Telematics and Informatics Reports

Show abstract

MAGIC: Malware behaviour analysis and impact quantification through signature co-occurrence and regression
2024, Computers and Security

Show abstract

BenchMFC: A benchmark dataset for trustworthy malware family classification under concept drift
2024, Computers and Security

Show abstract

Harnessing the advances of MEDA to optimize multi-PUF for enhancing IP security of biochips
2024, Journal of King Saud University - Computer and Information Sciences

Show abstract

## A comprehensive analysis combining structural features for detection of new ransomware families

2024, Journal of Information Security and Applications

Show abstract

## A survey on machine learning techniques applied to source code

2024, Journal of Systems and Software

Show abstract

View all citing articles on Scopus

**Daniele Ucci** is a Ph.D. student in Engineering in Computer Science at Department of Computer, Control, and Management Engineering “Antonio Ruberti” at Sapienza University of Rome. He received the master degree with honors in Computer Engineering in Computer Science in 2014 A.Y.. His research interests mainly focus on Big Data and information security and privacy, with special regard to malware analysis. During his master thesis, he has investigated topics related to business intelligence and Big Data. Currently, he is working both on privacy-preserving data sharing of sensitive information in collaborative environments and malware analysis based on machine learning techniques.

**Leonardo Aniello** is a Lecturer in Cyber Security at the University of Southampton, where he is also a member of the Cyber Security Research Group. He obtained a Ph.D. in Engineering in Computer Science in 2014 from “La Sapienza” University of Rome, with a thesis about techniques for processing Big Data in large-scale environments by adopting a collaborative approach, and with the aim of improving the timeliness of the elaboration. His research studies are currently focused on cyber security aspects, including malware analysis, blockchain-based systems and privacy-preserving data sharing. Leonardo is author of more than 30 papers about these topics, published on international conferences, workshops, journals and books.

**Roberto Baldoni** is a full professor at the Sapienza University of Rome. He conducts research (from theory to practice) in the fields of distributed, pervasive and p2p computing, middleware platforms and information systems infrastructure with a specific emphasis on dependability and security aspects. Roberto Baldoni is director of the Sapienza Research Center for Cyber Intelligence and Information Security and, at national level, is director of the Cyber Security National Laboratory. Recently, he has been appointed as coordinator of the National Committee for Cybersecurity Research born on February 2017 as an agreement between the Italian National Research Council and the Cyber Security National Laboratory. A partial list of his publications can be found at DBLP, at Scholar Google and at MIDLAB publication repository.

View full text