Machine Learning for Malware Detection: Beyond Accuracy Rates

Lucas Galante, Marcus Botacin, André Grégio, Paulo Lício de Geus

SBSEG 2019

Agenda

Agenda

Malware Increase

230K+ computer users hit by malware in Q2 2019: Report

DH Web Desk, Bengaluru, AUG 22 2019, 19:22PM IST | UPDATED: AUG 22 2019, 21:02PM IST

Figure: Increase of 46% in malware activity in Q2 of 2019.
https://tinyurl.com/y6qzn83h

Malware Classification

Anti-Malware Market Reviews: Industry Share, Trends, Analysis And Future Predictions For 2027

posted on AUGUST 23, 2019

At present, the vulnerability of computing devices and IT systems due to technical complexities, network security loopholes, and human errors are necessitating anti-malware applications. Antimalware applications protect computing systems against many types of malware such as viruses, ransomware, and spyware.

In common parlance, anti-malware is interchanged with anti-virus. However, the scope and capabilities of the former are wider not offered by anti-virus applications.

Figure: Necessity of antimalware applications. `https://tinyurl.com/y2bz5k58`

Agenda

Extracted Features

Table: Malware Features classified according extraction method (static and dynamic) and representation (discrete or continuous).

Static				Dynamic	
Discrete		Continuous		Both	
Embedded files	Dissasembly fail	Size sections	# headers	fork syscall	/proc access
/home string	ptrace syscall	/home string	# .dynamic	ptrace syscall	/home access
/sys string	Network strings	/sys string	# sections	socket syscall	passwd access
Linkage	Header present	passwd string	# symbols	mmap syscal	permission denied
UPX	passwd string	# libs	# relocations	SIGTERM	
fork syscall	compiler string	Size sample	# debug section	SIGSEGV	

Classification Overview

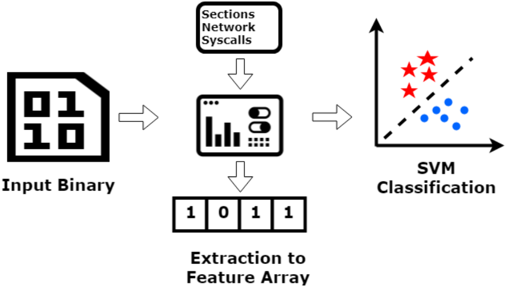

Figure: Overview of classification process.

Agenda

Importance of a Good Feature Extraction Procedure

SVM classification of static continuous features.

Kernel/Iter(#)	1000	10000	100000
Poly	49.32%	49.74%	49.95%
Linear	73.87%	77.64%	80.94%
rbf	84.92%	**84.92%**	84.92%

SVM classification of dynamic continuous features.

Kernel/ Iter (#)	1000	10000	100000
Poly	49.92%	49.76%	50.71%
Linear	93.73%	86.51%	86.73%
rbf	92.63%	**92.63%**	92.63%

Importance of Evaluated Datasets

Mixed dataset. Random Forest classification of static continuous features.

Max Depth/ Estimators (#)	16	32	64
8	**99.17%**	99.06%	99.20%
16	99.13%	99.06%	99.09%
32	99.09%	99.13%	99.17%

VirusTotal dataset. Random Forest classification of static continuous features.

Max Depth/ Estimators (#)	16	32	64
8	94.29%	**94.35%**	94.24%
16	94.24%	94.14%	94.08%
32	94.08%	94.14%	94.19%

Analyst Importance

SVM classification of dynamic continuous features.

Kernel/ Iter (#)	1000	10000	100000
Poly	50.91%	54.05%	58.16%
Linear	97.97%	97.56%	80.35%
rbf	98.54%	**98.54%**	98.54%

SVM classification of dynamic discrete features.

Kernel/ Iter (#)	1000	10000	100000
Poly	79.68%	79.91%	79.91%
Linear	96.48%	**96.48%**	96.48%
rbf	96.35%	96.35%	96.35%

What ML results teach us

Static feature importance

Static			
Discrete		Continuous	
Network strings	40%	Binary size	27%
UPX present	17%	# headers	16.70%
passwd strings	1.40%	# debug sections	0.20%

Dynamic feature importance

Dynamic			
Discrete		Continuous	
mmap	50%	# mmap	68%
fork	6%	# fork	10.80%
SIGSEGV	10.60%	# SIGSEGV	1.30%

Agenda

Conclusion

Our results show that:

- Dynamic features outperforms static features
- Discrete features present smaller accuracy variance
- Dataset's distinct characteristics impose challenges to ML models
- Feature analysis can be used as feedback information

Acknowledgement

This work is supported by:

- Brazilian National Counsel of Technological and Scientific Development
- CESeg assistance

Questions, Critics and Suggestions.

Contact

- **galante@lasca.ic.unicamp.br**

Complete version

- `https://github.com/marcusbotacin/ELF.Classifier`

Previous work

- `https://github.com/marcusbotacin/Linux.Malware`

Reverse Engineering Workshop

- Thursday @ 13:30