HOW I MET YOUR MALWARE

Machine learning driven malware detection

Giarduz Andrea - Grosso Veronica - Nannini Riccardo

OUTLINE

- Problem statement
- Machine Learning techniques
  - KNN
  - SVM
  - Logistic Regression
- Results and comparison

1

PROBLEM STATEMENT

Malware detection - a classification problem

MALWARE DETECTION

The aim: **CLASSIFY** binaries - whether they are *malicious* or *benign*

Breakdown of features obtained by **static analysis**

WHY MACHINE LEARNING?

Antivirus limits

Normal signature-based antimalware (or antivirus) suffer from stealth technique like polymorphism or metamorphism

Impossibility of perfect virus detector

The perfect antivirus does not exist, the virus detection problem is a variant of the halting problem

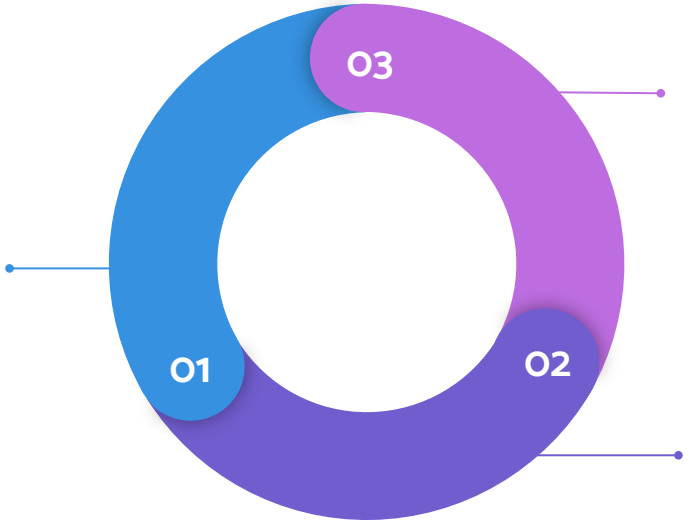
ML / Behavioral detection

In the last few years, antimalware softwares started focusing on different detection techniques, some of them based on machine learning

DATA PREPROCESSING

Standardization of the data

The data has been standardized following a Gaussian distribution with means = 0 and variance = 1.



Identification of the cross validation folds

5 fold cross validation is performed in order to tune the hyperparameters and find the optimal value to adopt.

Identification of training and testing data

80% of the dataset has been used exclusively for training purposes, whereas 20% only for testing.

2

MACHINE LEARNING TECHNIQUES

KNN, SVM, Logistic Regression

K-Nearest Neighbors

Cross-validation accuracy values with different hyperparameters

97.55%	97.56%	97.36%	97.17%	97.04%
1 NN	3 NN	5 NN	7 NN	9 NN

3 NN

Optimal value

K-Nearest Neighbors

Cross-validation accuracy values with different hyperparameters

97.55%	97.56%	97.36%	97.17%	97.04%
1 NN	3 NN	5 NN	7 NN	9 NN

3 NN

Optimal value

KNN TEST DATA

- Prediction accuracy:

  97.60%

- Inference time:

  1038 ms

Support Vector Machine

Cross-validation accuracy values with different hyperparameters

76.33%	83.50%	90.20%	95.10%	96.02%
C=0.01	C=0.1	C=1.0	C=5.0	C=10.0

C = 10

Optimal value

Polynomial Kernel

Support Vector Machine

Cross-validation accuracy values with different hyperparameters

76.33%	83.50%	90.20%	95.10%	96.02%
C=0.01	C=0.1	C=1.0	C=5.0	C=10.0

C = 10

Optimal value

Polynomial Kernel

SVM TEST DATA

- Prediction accuracy:

  95.84%

- Inference time:

  353 ms

Logistic Regression

$\lambda = 1/C$

Cross-validation accuracy values with different hyperparameters

95.09%	**95.85%**	**95.96%**	**95.77%**
C=0.001	C=0.01	C=0.1	C=1.0
95.82%	**95.92%**	**96.10%**	**96.18%**
C=10.0	C=100.0	C=1'000.0	C=10'000.0

C= 10'000

Optimal value

Logistic Regression

$\lambda = 1/C$

Cross-validation accuracy values with different hyperparameters

95.09%	95.85%	95.96%	95.77%
C=0.001	C=0.01	C=0.1	C=1.0
95.82%	95.92%	96.10%	96.18%
C=10.0	C=100.0	C=1'000.0	C=10'000.0

C= 10'000

Optimal value

LOGISTIC REGRESSION TEST DATA

- Prediction accuracy:

  96.30%

- Inference time:

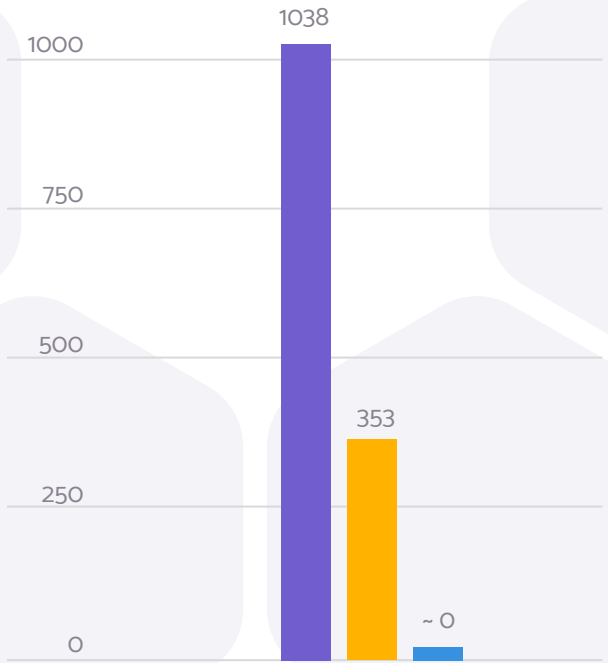
  ~ 0 ms

3

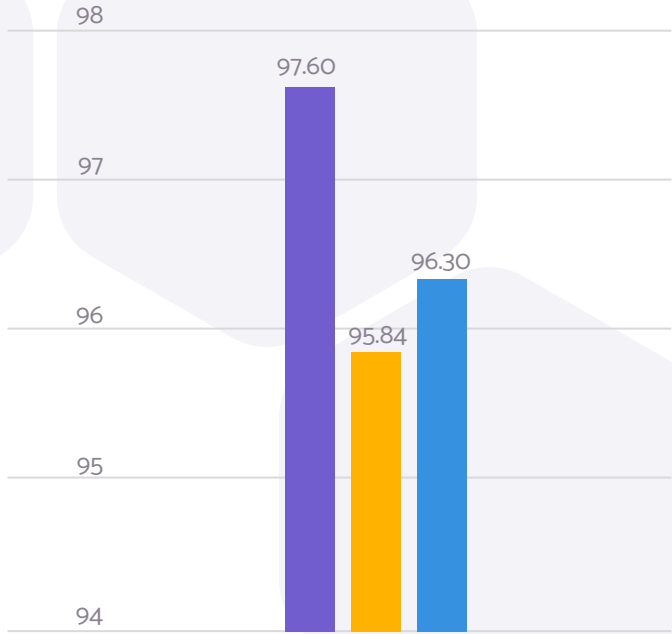
COMPARISON & CONCLUSION

Accuracy and inference time comparison

Inference time (ms)



Prediction accuracy (%)



KNN

SVM

Logistic regression

Thanks!

Machine learning driven malware detection final project

HOW I MET YOUR MALWARE

Giarduz Andrea - Grosso Veronica - Nannini Riccardo