KLASIFIKACIJA ZLONAMJERNOG SOFTVERA

Travanj 2020.

Stanišić Matea
Škrabo Petra
Terzanović Mateja
Tolja Margarita

Opis problema

Microsoft Malware Classification Challenge (*Kaggle*)

Dataset

- Više od 20 000 primjeraka *malware*-a
- Neraspakirani podaci
- Podatci za treniranje i testiranje
- 9 klasa
- .bytes i .asm datoteke
- Izvlačenje značajki

Opis problema

Cilj

- Izraditi model koji će što bolje klasificirati *malware*-e u odgovarajuće skupine
- Uspostaviti ravnotežu između kompleksnosti i performanse

Metodologija i plan istraživanja

Metodologija i plan istraživanja

- Odabir značajki (unvarijantna metoda, random forest metoda...)
- Konstruirati model i eksperimentirati s različitim skupovima značajki
- XGBoost, Bagging...
- *Jupyter* bilježnica i `Python`-ove biblioteke za strojno učenje

Mjera uspješnosti

- Točnost i *logloss*
- Provjera rješenja učitavanjem na *Kaggle*

Literatura I


Ahmadi, Mansour et al. (Mar. 2016). "Novel Feature Extraction, Selection and Fusion for Effective Malware Family Classification". In: DOI: 10.1145/2857705.2857713.


Harlick texture features. http://murphylab.web.cmu.edu/publications/boland/boland_node26.html. Dohvaćeno: 22-04-2020.


Hrvatska enciklopedija, entropija. https://www.enciklopedija.hr/natuknica.aspx?ID=18042. Dohvaćeno: 22-04-2020.


Hrvatski mrežni riječnik. http://ihjj.hr/mreznik/page/pojmovnik/6/. Dohvaćeno: 22-04-2020.


Making Sense of Logarithmic Loss. https://datawookie.netlify.app/blog/2015/12/making-sense-of-logarithmic-loss/. Dohvaćeno: 22-04-2020.


Microsoft malware classification Challenge. http://arxiv.org/abs/1802.10135. Dohvaćeno: 13-04-2020.

Literatura II

- *Microsoft malware classification Kaggle Challenge*. https://www.kaggle.com/c/malware-classification. Dohvaćeno: 13-04-2020.
- Nataraj, Lakshmanan et al. (July 2011). "Malware Images: Visualization and Automatic Classification". In: DOI: 10.1145/2016904.2016908.
- *Tomislav Lipić i Matija Piškorec, Vježbe kolegija Strojno učenje*. https://github.com/pmf-strojnoucenje/Vjezbe. Dohvaćeno: 22-04-2020.
- *Tomislav Šmuc, Predavanja iz kolegija Strojno učenje*. Dohvaćeno: 22-04-2020.
- *Understanding and Detecting Malware Threats Based on File Size*. https://d3gpjj9d20n0p3.cloudfront.net/fortiguard/research/DetectingMalwareThreats.pdf. Dohvaćeno: 22-04-2020.
- *What Can N-Grams Learn for Malware Detection?* https://www.edwardraff.com/publications/what_can_ngrams_learn.pdf. Dohvaćeno: 22-04-2020.