Klasifikácia a podobnosť rodín malvéru

Peter Chomič

1Im, 2018 - 2019

Vedúci práce: JUDr. RNDr. Pavol Sokol, PhD.

Konzultant: Mgr. Ladislav Bačo

Autor: Bc. Peter Chomič

Ciele:

1. Vytvorenie dátovej sady pre klasifikáciu vzoriek malvéru
2. Porovnanie prístupov ku klasifikácii a určovaniu podobnosti malvéru
3. Vytvorenie modelu, implementácia a vyhodnotenie pre klasifikáciu vzoriek malvéru do jednotlivých rodín malvéru.

1 Popis cieľov

Cieľom práce je porovnať rôzne skupiny klasifikátorov pri budovaní skupiny klasifikátorov (ensemble). Okrem toho vyskúšame viaceré prístupy k budovaniu tejto skupiny. Prínos práce spočíva v tom, že základné klasifikátory nedostanú rovnaké dáta, ale každý dostane inú podmnožinu dát získaných zo statickej a dynamickej analýzy. Takto pokryjeme celú škálu charakteristík a správania sa danej vzorky bez toho, aby sme mali vstupné dáta s príliš veľkou dimenzionalitou. Týmto ušetríme čas pri klasifikácii a zabránime, aby nejaký klasifikátor vykazoval „kliatbu dimenzionality“. Potom pre rôzne skupiny klasifikátorov porovnáme náš postup s postupom, ktorý využíva jednotné mnohodimenzionálne dáta z pohľadu presnosti, času a výkonu.

Samotné ciele práce korešpondujú s postupom. Najprv je potrebné získať dáta zo statickej a dynamickej analýzy. Tieto dáta môžu zahrňovať samotné spustiteľné súbory, metadáta, API, systémové volania (system calls), linkované knižnice, sekvencie inštrukcií, samotné zdrojové kódy, závislosti (dependencies) a iné.

Z týchto dát je potom potrebné získať charakteristiky (features), ktoré ich popisujú. Medzi ne patrí výskyt n-gramov, ich frekvencie, grafy volaní, prípadne Markove reťazce (Markov chains) ich priebehu. Zo spustiteľných súborov možno urobiť obrázky pre konvolučnú sieť či histogram entropie.

Charakteristík bude však veľa aj v prípade, že každý klasifikátor dostane len ich podmnožinu. Je však možné použiť algoritmy na výber podmnožín charakteristík ktoré vyberú také charakteristiky ktoré pokrývajú informáciu o triede. Navyše eliminujú ostatné, nadbytočné (feature selection). Medzi tieto algoritmy patrí genetický algoritmus či simulované žíhanie (simulated annealing). Všeobecne sa tieto metódy delia na filter, wrapper a embedded. Pre rôzne klasifikátory sú potrebné rôzne

algoritmy. V tomto kroku už musíme mať zvážené, ktoré klasifikátory použijeme, aby sme pre nich mohli nájsť správne podmnožiny charakteristík.

V ďalšom kroku je potrebné vyjadriť charakteristiky v takom tvare, aby mohli slúžiť ako vstupné dáta pre klasifikátory. Väčšina klasifikátorov prijíma na vstup normalizované vektory (na interval <0,1>). Konvolučná neurónová sieť (CNN) je schopná spracovať obrázky, a pri algoritme K-najbližších susedov (K-nearest neighbours) záleží od definície vzdialenosti.

Keďže tieto vektory môžu mať veľmi vysokú diemnzionalitu, je dobré použiť algoritmy na jej zníženie (feature dimensonality reduction). Príkladom je autoencoder pri neurónových sieťach alebo Kernel PCA pre support vector machine (SVM). Medzi iné metódy patrí locality sensitive hashing a random projection.

Hlavným krokom klasifikácie je použitie samotných klasifikátorov. Pri klasifikácii malwaru do rodín sa používajú rôzne typy neurónových sietí (hlboké - CNN, rekurentné aj dopredné typy), potom SVM, rozhodovacie stromy (decision trees,DT), random forest (RF), multinomiálna logistická regresia, naive bayes, ale aj evolučné algoritmy, algoritmy pre podobnosti grafov (volaní) a aj algoritmy z bioinformatiky pre klasifikáciu génových sekvencií. S architektúrou klasifikátorov možno tiež experimentovať pre dosiahnutie lepších výsledkov.

Posledný krok procesu je vyhodnotenie. Aby sa získali lepšie dáta pre hodnotenie tak namiesto toho, aby sa vždy učilo na rovnakej vzorke sa používa k-fold cross-validation kde sa rozdelia na „k“ menších vzoriek a postupne sa k-1 použije na učenie. Pre samotné vyhodnotenie existuje viacero štatistických veličín, na ktoré sa možno zamerať. Medzi ne patrí presnosť (accuracy) – suma true positive a true negative sa delí veľkosťou vzorky, potom recall – true positive / (true positive + false negative) a precision – true positive / (true positive + false positive). Často sa používa F1 skóre, čo je harmonický priemer z precision a recall.

Literatúra

1. Saxe, J., Sanders, H.: Malware data science - attack detection and attribution, San Francisco, No starch press. 2018.

2. Monnappa, K.A.: Learning malware analysis, Packt. 2018.

3. Gibert, Daniel, Carles Mateu, and Jordi Planes. "An End-to-End Deep Learning Architecture for Classification of Malware's Binary Content." International Conference on Artificial Neural Networks. Springer, Cham, 2018.Dostupné na [1.12.2018]: https://link.springer.com/chapter/10.1007/978-3-030-01424-7_38

4. Yan, Jinpei, Yong Qi, and Qifan Rao. "Detecting malware with an ensemble method based on deep neural network." Security and Communication Networks 2018 (2018). Dostupné na [1.12.2018]:https://www.hindawi.com/journals/scn/2018/7247095/abs/

5. Yakura, Hiromu, et al. "Malware Analysis of Imaged Binary Samples by Convolutional Neural Network with Attention Mechanism." Proceedings of the Eighth ACM Conference on Data and Application Security and Privacy. ACM, 2018.

Dostupné na [1.12.2018]: https://dl.acm.org/citation.cfm?id=3176335

6. Jung, Byungho, Taeguen Kim, and Eul Gyu Im. "Malware classification using byte sequence information." Proceedings of the 2018 Conference on Research in Adaptive and Convergent Systems. ACM, 2018.
Dostupné na [1.12.2018]: https://dl.acm.org/citation.cfm?id=3264775

7. Gibert, Daniel, et al. "Classification of Malware by Using Structural Entropy on Convolutional Neural Networks." AAAI. 2018.

8. Ronen, Royi, et al. "Microsoft Malware Classification Challenge." arXiv preprint arXiv:1802.10135 (2018).
Dostupné na [1.12.2018]: https://arxiv.org/abs/1802.10135

9. Gibert, Daniel, et al. "Using convolutional neural networks for classification of malware represented as images." Journal of Computer Virology and Hacking Techniques (2018): 1-14.
Dostupné na [1.12.2018]: https://link.springer.com/article/10.1007/s11416-018-0323-0