Malware Detection

Hivan J. Sabr

Outline

- Introduction
- Problem
- Malware Detection Process
- Malware Detection Approaches
- Malware Detection Datasets
- Malware Detection Evaluation
- Conclusion

Introduction

- In recent years, almost every member of the society has been using the Internet for daily life
  - Social interactions
  - Online banking
  - Health related transaction
  - Marketing
  - ….
- Any software which intentionally executes malicious payloads on victim machines (computers, smart phones, computer networks, etc.) is considered as malware

Introduction (Continue.)

- There many type of Malware such as
  - Virus
  - Warms
  - Trojan Horse
  - Rootkit
  - ….
- To protect legitimate users and companies from malware, malware need to be detected
- Malware detection is the process of determining whether a given program has malicious intent or not.

Problem in Malware Detection

- The studies shown that the problem of detecting the malware is NP-complete.
- Virus detector for certain virus strain can be used to solve the satisfiability problem. Since satisfiability problem m is known to be NP-complete, so the detection of the malware is **NP-complete**.
- Therefore it is impossible to design an algorithm which can detect all malware.

Problem in Malware Detection

- Practically the new generation of malware use techniques such as encryption, oligomorphic, polymorphic, metamorphic, stealth, and packing methods to make detection process more difficult.
- This kind of malware can easily bypass protection software that is running in kernel mode such as firewalls, antivirus software.
- This makes practically almost impossible to detect all malware with single detection approach because the computational complexity of malware is not clear, and the detection of malware problem is proved to be **NP-complete**.

Malware Detections Process

- Data Gathering
- Static Analysis
- Dynamic Analysis
- Future Extraction
- Classification

Malware Detection Approaches

Approaches

- Signature-based
- Behavior-based
- Heuristic-based
- Model checking-based
- Deep learning-based
- Cloud-based
- Mobile-based
- IoT-based

Malware Datasets

- NSL-KDD dataset (2009)
    - https://www.kaggle.com/datasets/hassan06/nslkdd
- Drebin dataset (2014)
    - https://drebin.mlsec.org/

Sample Dataset Explanation

1. **Basic Features (1-9)**:
   - **1. duration**: Length of the connection in seconds.
   - **2. protocol_type**: Type of protocol (e.g., TCP, UDP, ICMP).
   - **3. service**: Network service (e.g., HTTP, FTP).
   - **4. flag**: Status of the connection (e.g., SF for normal).
   - **5. src_bytes**: Number of data bytes sent from source to destination.
   - **6. dst_bytes**: Number of data bytes sent from destination to source.
   - **7. land**: 1 if the connection is from/to the same host/port; 0 otherwise.
   - **8. wrong_fragment**: Number of wrong fragments.
   - **9. urgent**: Number of urgent packets.

Content Features (10-22):

- **10. hot**: Number of "hot" indicators.
- **11. num_failed_logins**: Number of failed login attempts.
- **12. logged_in**: 1 if successfully logged in; 0 otherwise.
- **13. num_compromised**: Number of compromised conditions.
- **14. root_shell**: 1 if root shell is obtained; 0 otherwise.
- **15. su_attempted**: 1 if "su" command is attempted; 0 otherwise.
- **16. num_root**: Number of root accesses.
- **17. num_file_creations**: Number of file creation operations.
- **18. num_shells**: Number of shell prompts.
- **19. num_access_files**: Number of access file operations.
- **20. num_outbound_cmds**: Number of outbound commands in an FTP session.

Malware Detection Evaluation

- The malware detections are evaluated in term of their accuracy to detect malware.
- Accuracy (Acc), Precision (Pr), Recall (Re), and F1 score are the four main classification metrics as follow:

$$\text{Acc} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}}$$

$$\text{Pr} = \frac{\text{TP}}{\text{TP} + \text{FP}}$$

$$\text{Re} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

$$\text{F1 score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}.$$

Conclusion

- Even though several new methods have been proposed by using these different malware detection approaches, no method could detect all new generation and sophisticated malware.
- The number, severity, sophistication of malware attacks, and cost of malware inflicts on the world economy have been increasing exponentially.
- Datamining and ML, new technologies such as deep learning, cloud, mobile devices, and IoT-based detection schemas have become popular.

References

- Aslan, Ö.A. and Samet, R., 2020. A comprehensive review on malware detection approaches. *IEEE Access*, *8*, pp.6249-6271.
- Hemalatha, J., Roseline, S.A., Geetha, S., Kadry, S. and Damaševičius, R., 2021. An efficient densenet-based deep learning model for malware detection. *Entropy*, *23*(3), p.344.