УДК 004.492.3

DEVELOPMENT OF A MACHINE LEARNING BASED METHOD FOR MALWARE DETECTION

Во С.А. (Университет ИТМО), **Воробьева А.А** (Университет ИТМО)
Научный руководитель – к.т.н., доцент Воробьева А.А.
Университет ИТМО

In this work, we present a model for the task of malware detection with machine learning, based on static analysis of the Portable Executable (PE) file. This work focusses on the feature selection step to achieve a good balance between the classification accuracy, the learning time and the storage required.

Введение. Each year, hundreds of millions of new malwares appear, making malware detection based purely on signature-matching impossible. Machine learning is a good countermeassure against modern malware, since it helps to detect malware without specific signatures. One of the biggest problems in Machine learning based malware detection is the time and storage required for a model to learn, since the number of features and the number of malware samples is large. Reducing the number of features can reduce the time and storage required, but it also decrease the accuracy achieved by the model. In this work, we evaluate the most-used groups of features in static analysis malware detection model and identify the best combination that achieve a good balance between detection accuracy and costs.

Основная часть. The Portable Executable (PE) file format is a file format for executables, object code, DLLs in Windows operating systems, encapsulates the information necessary for Windows OS to manage executable code. This includes dynamic library references for linking, API export and import tables, resource management data, thread-local storage data. This is a vital information in the sphere of static analysis malware detection. The most-used group of features in the task of static analysis for malware detection based on PE files are: general file information, header information, imported functions, exported functions, section information, byte histogram, byte-entropy histogram and string information. This work focusses on finding the suitable classification algorithm and the best combination among those groups of features. A dataset of 100 000 PE files of executable files is created, which includes 50 000 malwares and 50 000 benign programs. A model is constructed using the Light Gradient Boosting Machine. By using all the features above, the model took over 25 hours to vectorize the dataset and 5 hours to train. With a threshold of 0.85, the model gives an accuracy of over 93%, with a FPR less than 0.6%. We also suggest some other combination of features, which give an acurracy of over 91%, FPR of less than 1%, with significantly lower time and storage required to train.

Выводы. In this work, we created a dataset of 100 000 PE files, built a Machine learning model based on Light Gradient Boosting Machine for the task of malware detection. Based on the training time, the storage required and the classification accuracy achieved, we present several good combinations of features groups, suitable for practical uses.

Во С.А. (автор) Подпись

Воробьева А.А. (научный руководитель) Подпись