Linux Malware Detection and Classification

Today, Linux is powering over 80% of smartphones, 99% of supercomputers, 97% of consumer electronics, 95% of web, ecommerce giants like Amazon, PayPal, Walmart, etc.

Adding to this, there has been a recent increase of use of Linux as mainstream desktop operating system.

But this shift doesn’t come without cons. With the increasing popularity of Linux, it is also witnessing increased development of Linux malwares at a rapid pace. And with this, we need to urgently address the security aspect of Linux.

In this paper, we have focussed on a Machine Learning based approach towards identifying and classifying the Linux malwares, embedded in the ELF file format.

ELF stands for Executable and Linkable file format, and is the standard binary file format for Linux systems.

The first step is Data Collection – in this, we have collected about 4500+ ELF files of all types from various sources. We have considered 6 main categories: Benign ELFs (ELFs that are completely safe and possess no threat), Backdoors, DDoS, Virus, Botnets and Trojans.

After data collection, we move on to Feature extraction – ELF files can be read using programs like readelf (available for Windows and Linux), and we used python module named pyelftools for our project. We extracted the features from file and stored them in CSV file for further processing. We worked upon about 200+ features for our analysis. We also labelled the data by adding a “type” column containing the category of that ELF (to be used as target by our model).

The data generated is not fit currently for processing. So, we had to do data cleaning which involved filling in missing values, and converting non numeric entries to numeric ones. After this step, we move on to Feature selection.

Of the 200 total features, not all of them contributed significantly towards helping in predicting the class of ELF. So, to improve the time required to train the model, and its accuracy, and also to reduce the model’s complexity, we did feature selection. All the 200+ features were ranked by their “importance” towards the target. Then we chose the best and significant features to train our model.

Now we had to select the algorithm to train our model. For this, we tried and tested our model on various classifiers available for supervised learning (namely Random Forest Classifier, Logistic Regression, Linear Support Vector Classification, and Multinomial Naive Bayes) and compared their performance. For this case, we found Random Forest classifier giving the best results.

So, we trained our model using random forest classifier using the train-test split method in a 70-30 ratio, and saved the trained model.

Our model is now ready to be used, and was found to be able to classify unknow ELF files with decent accuracy.

Scope of work and Further improvements – this paper focusses on cybersecurity (particularly of Linux systems), which is vital. And use of Machine Learning can help us greatly in identifying useful insights of the malwares. Further improvements in the paper include the addition of ability of model to constantly evolve.