Android Malware Classification

Aya Mourad, Zoulfikar Shmayssani

May 2020

Methodology

In this project, we suggest feature-based Android malware detection, including standard permissions, intents, and APIs. Permissions and Intents features are formed by extracting standards - starts with "android." - features used in the Android system. We got a total of 272 and 203 features, respectively. APIs features are obtained by extracting standard APIs - starts with "android." - from up to 50 APIs per APK. We got a total of 2691 features. We aim to improve the detection accuracy by combining permissions, intents, and APIs sets, obtaining a total of 3166 features. As a result, every APK can be represented as a binary vector, i.e., V where $V_i = 1$ if and only if the APK has the i^{th} feature and $V_i = 0$ if corresponding APK does not indicate the feature.

Classification

We have done feature selection using SelectFromModel of importance threshold 0.001, where we obtain 356 features. We perform Logistic Regression classification on the selected features achieving training accuracy 97.98%, testing accuracy 95.1%, over/underfitting of the model 2.9%, and log loss 0.14. We also used Neural Network Model - Keras without feature selection consisting of five layers with sigmoid as an activation function and SGD with 0.04 learning rate achieving training accuracy 97.85%, testing accuracy 96.02%, over/underfitting of the model 1.83%, and log loss 0.14.

References

[1] Peiravian, N., Zhu, X. (2013, November). Machine learning for android malware detection using permission and api calls. In 2013 IEEE 25th international conference on tools with artificial intelligence (pp. 300-305). IEEE.