Comparing Performance of Malware Classification on Automated Stacking

Yu-Wen Chen

Department of Computer Systems Technology

18th Annual City Tech Poster Session
November 19, 2020

Abstract

Stacking in machine learning allows multiple classification or regression algorithms to work together with a goal to enhance performance. To understand the risky properties of malware contamination in a system, it is important to accurately classify malware type first. Malware classification is the procedure of labeling the families of malware. In this work, we automate stacking with 7 machine learning algorithms and 3 boosting algorithms. The experimental results show a 99.2% accuracy is achieved from a multilayer perceptron network with AdaBoost classifier, which outperforms other models on the malware API call dataset.

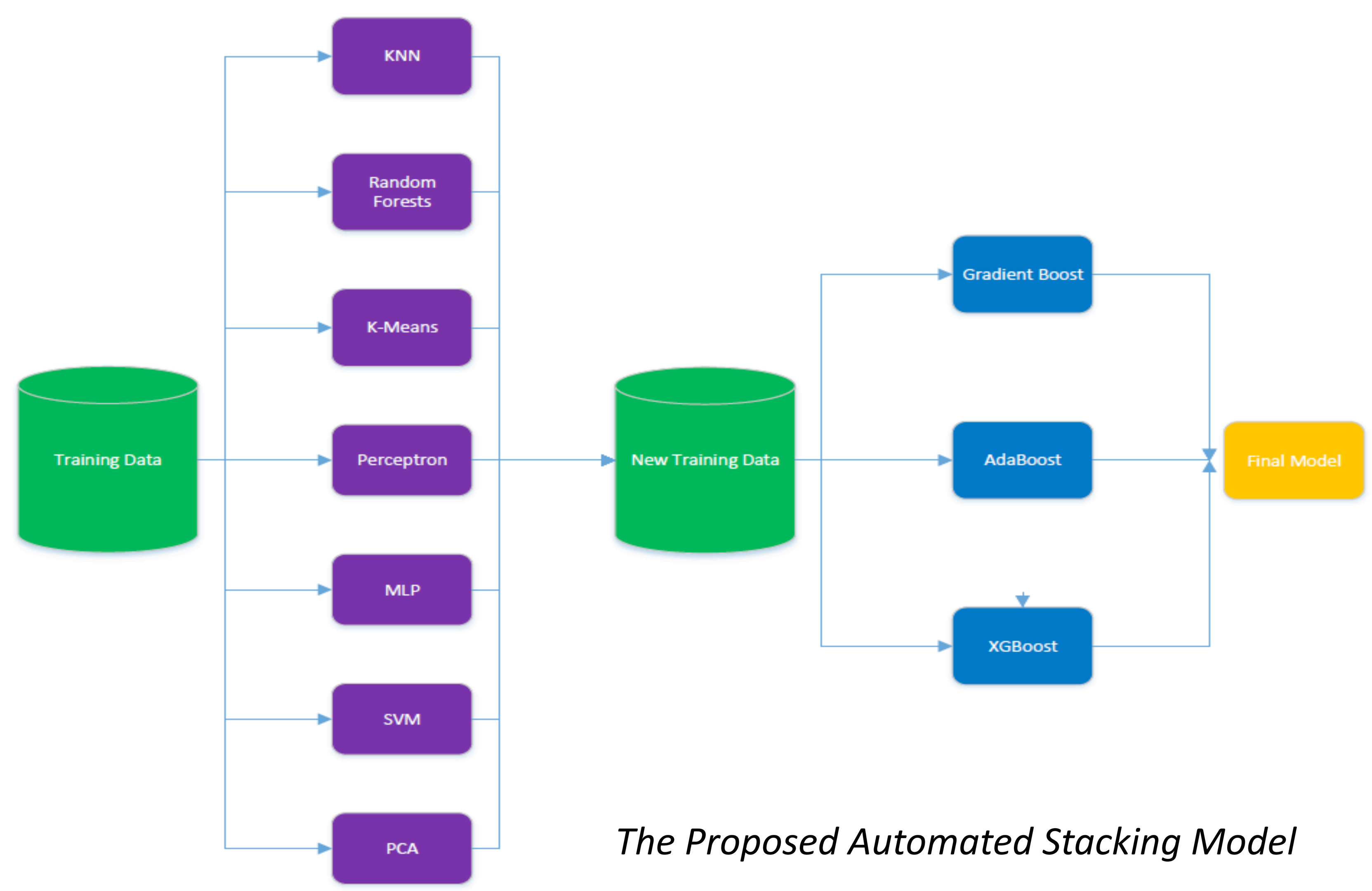
Introduction

The exponential growth of malware has created a significant threat in our daily lives, which heavily rely on computers running all kinds of software. Sophisticated malware, such as Dugu 2.0 exploiting a number of zero-day vulnerabilities, is hard to be detected using traditional virus scanning approaches. Signature-evading, polymorphic viruses such as the Storm worm, which proliferates tens of thousands of variants monthly, poses a challenge to antivirus software based on static signatures. The number of malware variants and complex malware turns malware analysis into a big data problem, a new challenge in the research community.

Big data analytics shares several advantages including learning from sample datasets to create a prediction model, which is used to detect new malware following a trend or a pattern learned from samples. Studies have shown some acceptable results. However, each model is fully dependent on a dataset and often overfitting occurs, which decreases model performance drastically. Stacking machine learning is a technique that ensemble some weaker models, and each of which is in charge of a part of the problem with a prediction. Multiple predictions are then forwarded to create a model for predicting the same target. In this work, we propose to apply stacking machine learning using boosting algorithms to automate the malware analysis to enhance model performance.

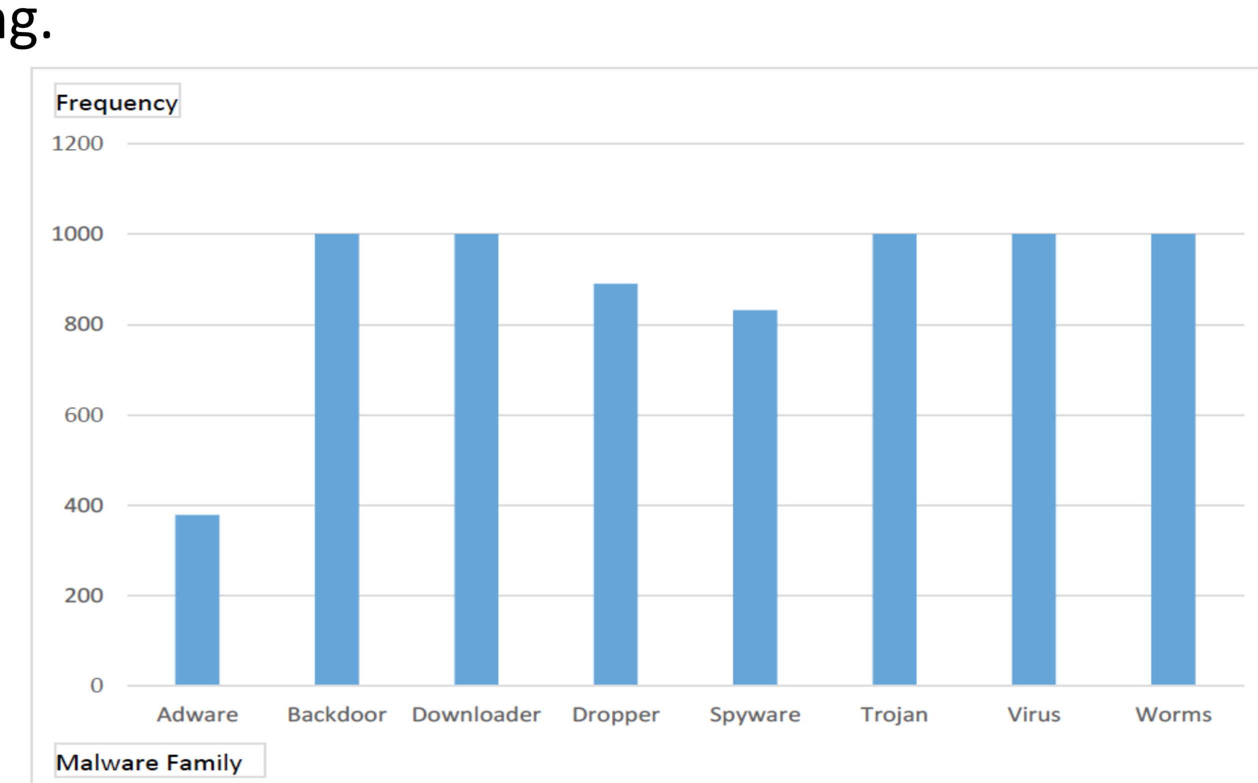
Methodology

We take the Malware API 2019 dataset from IEEE Data Port. Then the dataset is fed into the data preprocessing. Second, each machine learning model is paired with 3 boosting algorithms to enhance performance. Finally, in the last layer, the best prediction is selected for the final prediction.



The Proposed Automated Stacking Model

Malware API Call Dataset: The Mal-API-2019 dataset is used and can be retrieved from IEEE DataPort. In the dataset, there are eight families of malware: Trojan, Backdoor, Downloader, Worms, Spyware Adware, Dropper, and Virus. The first 1500 malware samples with 55 attributes are used in our study. For each malware sample, the first 100 API calls are sliced for analysis. In our experimental setting.



Malware Family Distribution in the Dataset Used in the Study

Data Preprocessing: There are several repeated API calls. The purpose of the experiment is to detect malware families. So we eliminate the repeated API calls. Then we implement the non-repeated API dataset to label encoder. Then the data is ready for machine learning algorithms.

Experiments and Results

We implemented Automated Stacking to the Malware API dataset. We applied seven different algorithms including KNN, Random forest, Kmeans, Perception, MLP, SVM, PCA to learn and three different boosting algorithms: XGboost, AdaBoost, Gradient Boosting to test the data. We split the dataset to 70 percent (training) and 30 percent (testing). We used the ensemble method to calculate the accuracy score. The proposed automated stacking machine learning model is implemented in Python to analyze the malware API call dataset. Table 1 shows the accuracies of the model at the last layer with three boosting algorithms XGBoost, Adaboost, and Gradient Boosting. The results indicate that the highest accuracy 99.2% occurs at the MLP-Adaboost ensemble method. The MLP classifier is a neural network-based, and in the experiment, the number of hidden layers is set to 100. AdaBoost is a boosting algorithm that covers the weakness of MLP classifier by adding more weights to weak learners. This combination further enhances performance.

KNN	Random Forest	K-means	Perceptron	Paired Algorihns
88.6%	70.4%	94.7%	55.8%	XGBoost
58.3%	58%	56.6%	23.3%	Adaboost
53.3%	75.6%	78.1%	98.2%	GradientBoost

MLP	SVM	PCA	Paired Algorihns
58.8%	34.1%	87.2%	XGBoost
99.2%	36.6%	89.2%	Adaboost
97.4%	98.1%	98.8%	GradientBoost

Table 1: Prediction Accuracy at the Last Layer in the Proposed Automated Stacking Machine Learning Model

Conclusion

Stacking machine learning provides an instrument to use different algorithms for training and testing data for different aspects. Boosting enhances weak learners by aggregating each of their strengths. In this work, we propose an automated stacking model that combines stacking and boosting to malware classifications. Our results show that the best performance coming from the MLP-Adaboost Classifier.

The work presented in this poster is from paper: Nusrat Asrafi, Dan Chia-Tien Lo, Reza M. Parizi, Yong Shi, and Yu-Wen Chen. Comparing Performance of Malware Classification on Automated Stacking. ACM Southeast Conference (ACMSE), April 2–4, 2020, Tampa, FL, USA.