Using Machine Learning to Detect Computer Network Security Threats

Jogalekar, Anushka

Purpose: The purpose of this project is to find out if/how machine learning can be used to detect network security threats. Hypothesis: Using machine learning tools in the statistical computation programming language R, it should be possible to train a program for detecting signatures of network security threats such as DOS, R2L, U2R and probing (supervised learning). Procedure: 1. Equipment: Dell laptop, RStudio, the R libraries gmodels, caret, c50, and e1071, and KDD cup internet traffic sample data published by MIT Lincoln Labs. 2) Training the program for normal and anomalous traffic: the R algorithms Naive Bayes and Decision Tree using the reduced KDD dataset (~500,000 packets). The training dataset has 41 comma-separated features. 3) Prediction: There were two major methods explored in this experiment: Naive Bayes and Decision Tree. Effectiveness was measured with confusion matrix.   Results: Both Naive Bayes and Decision Tree algorithms were able to detect all 23 types of network intrusions. Naive Bayes had 80 percent net accuracy (true positives and true negatives). Decision Tree had 99% net accuracy, observed in three trials each. Running Naive Bayes took 1.45 hours on average for 5e+6 packets. Running Decision Tree took 7.25 minutes on average per trial. Conclusion: “R” language ecosystem has powerful data processing and machine learning modules that can be effectively used for detecting internet security threats. Decision Tree proved to be a very effective algorithm for network security, with 99 percent accuracy, and can be effective in real time network monitoring. Naive Bayes can be a second choice for simplicity, but not suited for real time detection of security threats. Results are easily extensible for IOT security using hybrid dataset.

Awards Won:

Oracle Academy: Award of $5,000 for outstanding project in the systems software category.