Countermeasures

Proposed By:
Rahul Alapati
Sutanu Bhattacharya
Suraj Raymond D'Souza

In the real world, we need to adopt a proactive rather than a reactive approach for defending against adversarial attacks on our Author Identification Systems (AIS). Proactive defenses aim to prevent future attacks.

1. <Exploratory, Targeted, Integrity> Attack

Exploratory attacks can be detected by analyzing the test set being used for the classification of the trained AIS. If a disproportionate amount of data is observed near the decision boundary of the classifier, then we can determine that our AIS is being probed. For example, in Assignment-4 "Probing via an EC", the main objective was to evolve a population of feature vectors by probing the classifier, such that the classifier will classify the feature vectors as belonging to exactly one, two or three authors. In this case, a lot of data will be present near the decision boundary and an attack can be signaled right away.

Targeted attacks like in Assignment-4 for evolving feature vectors for exactly one, two or three authors, are more sensitive to variations in the decision boundary, because boundary movement is more likely to change the classification of the relevant points which in turn can lead to an integrity attack.

By detecting the attacks on our system, we will be able to know about adversary's capabilities and use that to develop countermeasures. In case of an Exploratory, Targeted, Integrity attack, the objective of the countermeasures should be to prevent the adversary from accessing the true decision boundary. To achieve that we can use information hiding and randomization as our countermeasures. We can make the classifier give incorrect feedback, by randomly changing the placement of the decision boundary. Also, we can place a theoretical bound on the information that can be retrieved by the adversary via probing.

2. <Causative, Indiscriminate, Privacy> Attack

Causative attacks can be detected by using a special test set. This set should consist of different feature vectors comprising of all the simulations of the known attacks. Once the classifier has been trained, misclassifying a large number of writing samples can indicate an attack.

The objective of the countermeasures should be to detect feature vectors that may cause harm to the classifier if trained on. One should try to identify the variance of each feature vector from the training set. Once detected harmful, all such feature vectors can be removed from the training set. To determine the impact of indiscriminate causative attacks, we can split the training data into different possible combinations of training and validation sets, then train the classifier on all possible subsets of training data and then test it on corresponding validation set. Then to identify the presence of malicious feature vectors, we can compare the performance of the classifier on different combinations of training and validation sets. The combination with disproportionate number of false positives and false negatives, can be further analyzed for the presence of malicious training samples.

Also, we can constrain the hypothesis space that the classifier considers, with the use of regularization. With regularization, we can restrict of the choice of hypothesis and remove all the complexities that an adversary might exploit. By developing countermeasures like these, we can also ensure the secrecy or privacy of the users.

3. <Exploratory, Indiscriminate, Integrity> Attack

Exploratory attacks can be detected by running a clustering algorithm against the data classified by the classifier. A sudden appearance of a large cluster near the decision boundary could indicate that the system is under attack and is being probed.

The objective of the countermeasures in case of an Exploratory, Indiscriminate and Integrity attack, should be to prevent the adversary from learning the true decision boundary. We can achieve this by hiding certain information from the adversary or by providing disinformation to the adversary. In case of information hiding, we can place a theoretical bound on the information, an adversary could gain by observing the behavior via probing. In case of disinformation, we can confuse the adversary by providing incorrect information. This role reversal of the learner and adversary, helps us in keeping a tab on the information seen by the adversary, thereby making it computationally intensive and difficult for the adversary to learn the decision boundary of the AIS. This in turn makes all the other attacks difficult.

References:

1. Barreno, Marco, et al. "Can machine learning be secure?." Proceedings of the 2006 ACM Symposium on Information, computer and communications security. ACM, 2006.
2. Udam Saini, “Machine Learning in the Presence of an Adversary: Attacking and Defending the SpamBayes Spam Filter” EECS-2008-62, University of California at Berkeley, 2008.
3. Biggio, Battista, and Fabio Roli. "Wild patterns: Ten years after the rise of adversarial machine learning." Pattern Recognition 84 (2018): 317-331.