Paper: *Subpopulation Data Poisoning Attacks* **.** *link*
Authors: *Matthew Jagielski, Paul Hand, and Alina Oprea, Northeastern University.*

Objective: To introduce a novel poisoning attack technique known as "Subpopulation attack" which is an incorporation/combination of availability (misclassifying as many data points as possible) and targeted (misclassifying targeted data points) attacks. Hence, the goal, in a sense, is to misclassify a certain subpopulation (called as Target) without misclassifying any other subpopulations in the training dataset.

1. **Assumptions and Knowledge:**
   - In a very diverse datasets, the performance of a sufficiently complex model on one subpopulation can become uncorrelated with the performance on a different subpopulation. The paper points out that the financial community might be vulnerable from this attack where financial datasets are large and diverse.

   - The proposed algorithm is a black-box in the target's dataset and model, where modification of data points at test time is not required. Black-box here means the attacker has no knowledge of the dataset or the model.

   - The adversary does not have access to the training dataset *D* but does have an auxiliary dataset *D auxiliary,* which is distinct from D but sampled from the same distribution.

   - Since the size of the attack is small relative to the overall dataset, the attack is stealthy.

2. **Objective of Poisoning Attacks:**
   a. Misclassify as many points as possible (availability attack)
   b. Misclassify a single target point (targeted attack)
   c. Misclassify out-of-distribution points (backdoor attacks)

By laying out these objectives, the paper asserts that the objective of their attack is different from the poisoning attacks done so far. Their objective is to target a certain subpopulation i.e misclassifying as many points as possible but only for the chosen subpopulation, putting their objective between (a) and (b), and hence stating as a novel attack.

3. **Definition:**
   - A subpopulation attack consists of a dataset of contaminants Dp and a filter function F : X -> {0, 1}. Dp is constructed to minimize the collateral damage (on points not in the subpopulation) and maximize the target damage (on points in the subpopulation) when appended to the training set

   - Goal: impact the predictions on inputs coming from a subpopulation in the data, but do not impact the performance of the model on points outside this subpopulation. The paper claims this is achieved by the filter function.

- Two metrics to measure the effectiveness of the attack:
    i) Collateral damage = E [for points not in the subpopulation (misclassified points in poisoned dataset - misclassified points in unpoisoned dataset)].
    Lower the value, better the attack.
    ii) Target damage = E [for points in the subpopulation (misclassified points in poisoned dataset - misclassified points in unpoisoned dataset)].
    Higher the value, better the attack.

4. Attacking Model:

a. **Filter Function Selection**:

FEATUREMATCH: If the attacker has a certain subpopulation in mind, feature match can be a good tool during the population selection process. For example: target subpopulation can be based on a certain gender, race etc.

$$F = (\text{”race”} = \text{”black”} \wedge \text{”gender”} = \text{”male”})$$

CLUSTERMATCH: The attacker can alternatively also use clusters to attack. Any clustering algorithm which generators meaningful data should work. “For example, consider an adversary who wishes to disrupt street sign detection in a self-driving car through a subpopulation attack, they could run clustering to identify vulnerable street signs which will be easiest to target, and attack those, increasing the impact and stealth of their attack.”

The paper later evaluates and compares both feature and cluster match and presents us with evidence and reason on which might be a better choice.

b. Flipping the labels

The adversary needs to pick an attack size, which should be comparable to the size of the subpopulation itself. To add contaminants, they sample points satisfying the filter function from D_{aux} and add these to the training set with flipped labels. This achieves this goal of misclassifying points satisfying the filter function.

Select a cluster from your pool and flip the label, which results: (x,y) -> (x,1-y). This paper selects these subpopulations arbitrarily, and uses a random flipping poisoning attack to generate the poison set.

Now, we have D_{p} i.e. the poisoned dataset and are ready for the attack. Hence, we train the model with D and D_{p} together which poisons the model.

5. Results: The paper shows that the CLUSTERMATCH is a better choice for selecting a filter function. The subpopulation attacks with CLUSTERMATCH achieves higher target damage thought they incur slight higher collateral damage than FEATUREMATCH. This further supports the hypothesis that clusters can be learned somewhat independently of the rest of the dataset.