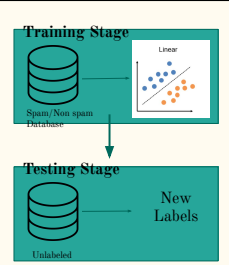
Explaining Vulnerabilities to Adversarial Machine Learning Through Visual Analytics

Yuxin Ma, Tiankai Xie, Jundong Li, Ross Maciejewski

Vulnerabilities in Machine Learning

Filtering Spam Emails

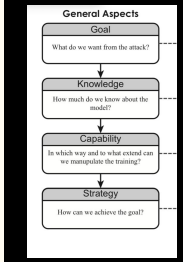


The Contribution of This Paper is:

- A visual analytics framework that supports the examination, creation, and exploration of adversarial machine learning attacks;
- A visual representation of model vulnerability that reveals the impact of adversarial attacks in terms of model performance, instance attributes, feature distributions, and local structures.

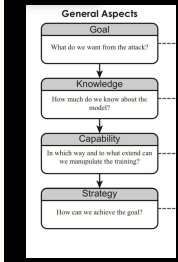
The four main features of an adversary are the adversary's ...

- **Goal**
  - Targeted attack
  - Reliability attack
- Knowledge
- Capability
  - Poisoning attack
  - Evasion attack
- Strategy



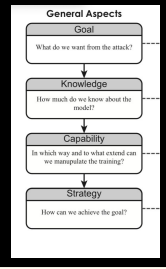
The four main features of an adversary are the adversary's ...

- Goal
  - Targeted attack
  - Reliability attack
- **Knowledge**
  - **Black box**
  - **White box**
- Capability
- Strategy



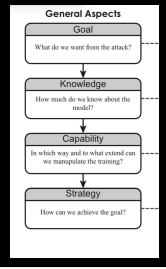
The four main features of an adversary are the adversary's ...

- Goal
  - Targeted attack
  - Reliability attack
- Knowledge
- **Capability**
  - **Poisoning attack**
  - **Evasion attack**
- Strategy



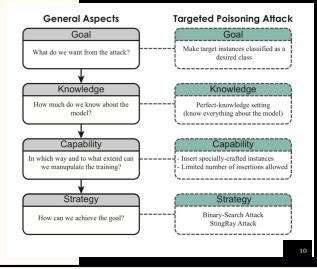
The four main features of an adversary are the adversary's ...

- Goal
  - Targeted attack
  - Reliability attack
- Knowledge
- Capability
  - Poisoning attack
  - Evasion attack
- **Strategy**



To demonstrate the proposed visual analytics framework, we focus our discussion on:

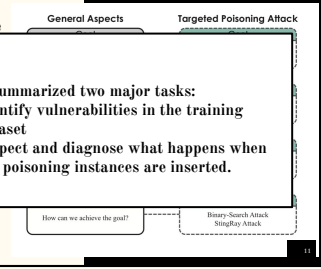
Targeted Data Poisoning Attack



To demonstrate the prop
anal
fram
focu
disc

Targ
Pois

- We have summarized two major tasks:
  1. Identify vulnerabilities in the training dataset
  2. Inspect and diagnose what happens when the poisoning instances are inserted.



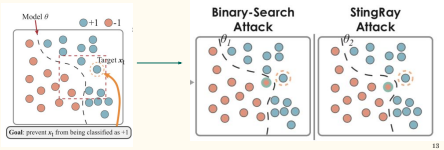
VISUAL ANALYTICS FRAMEWORK

The framework supports three main activities:

1. **Vulnerability analysis**
2. **Attack space analysis**
3. **Attack results analysis**

Vulnerability Analysis

- **Core idea: To change the label of the target instance**
  - **Attack algorithms: Binary-Search Attack & StingRay Attack**



Vulnerability Analysis

- **Vulnerability Measures (to explore the potential weaknesses in the model):**
  - **Decision Boundary Distances (DBD)**
  - **Minimum Cost for a Successful Attack (MCSA)**

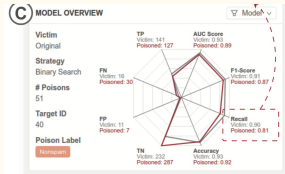
Visualizing the Attack Spac.

- **Vulnerability Measures:**
  - **Decision Boundary Distances (DBD)**
  - **Minimum Cost for a Successful Attack (MCSA)**
  - **Performance metries of the poisoned model (Accuracy, Recall, etc.)**

- **Each instance in the training dataset is measured based on these vulnerability measures.**
- **Video**

Attack Detail Analysis

Model Overview



Radar Chart for Model Performances

- TN
- FN
- TP
- FP

- Acc
- Recall
- F1
- ROC

Attack Detail Analysis

T-SNE Projection
- Overview of the data distribution

Instance View
- Instances Attribute
  - Details of Instances

- Key Attributes for Instances:
  - Decision Boundary Distances
  - Classification Probabilities
  - Labels of K-NNs

Attack Detail Analysis

T-SNE Projection
Overview of the data distribution

Instance View
- Instances Attribute
  - Details of Instances

- Key Attributes for Instances:
  - Decision Boundary Distances
  - Classification Probabilities
  - Labels of K-NNs

Length: Classification Probability
Positive Class
Negative Class
Arrow: from the prediction of the victim model to the poisoned one
Virtual Decision Boundary
Decision Boundary Distance
Positive
On the decision boundary
Negative

Attack Detail Analysis

Feature View

- Data Distributions on Features
  - Instances in the spam/non spam classes
  - Poisoning Instances
- Feature Importance Rankings
  - In the victim model
  - In the poisoned model
  - Differences

Attack Detail Analysis
Local Impact View

- The target would be influenced either by its neighbors or the poisoned instances

KNN Graph Building

Find kNNs
Build kNN Graph
Any Poisoning Instance

(a) kNN Graph Building

Attack Detail Analysis
Local Impact View

- The target would be influenced either by its neighbors or the poisoned instances

Innocent Instances
Target Instance
Poisoning Instance

KNN Graph Building

Find kNNs
Build kNN Graph
Any Poisoning Instance

(a) kNN Graph Building

Case Study

Critiques

- Strengths
  - Two stage design in the interface
  - Very user friendly
  - Multi Faceted Analysis
- Weaknesses
  - Only allows to attack one instance
  - Speed on bigger datasets and more complicated models
  - Scalability (Visual design, Attack Algorithm).
    - Case studies are too simplified.

1 Data Table View
2 The Selected Target
3 Model Overview
4 Projection View
5 Local Impact View
6 Instance Attribute View

Thank you!
QA

Attack Detail Analysis
Local Impact View

Innocent Instances

Target and Innocent Instances

Inner Ring: The class distribution of kNNs in the **victim model**
Outer Ring: The class distribution of kNNs in the **poisoned model**

Color: The predicted label
Texture: Whether the predicted label is flipped from the victim model
Size: Classification Probability