CERIAS

the center for education and research in information assurance and security

A Game Theoretic Framework for Adversarial Learning

Murat Kantarcioglu, Bowei Xi, and Chris Clifton

Introduction

- Many adversarial learning problems in practice
  - Intrusion Detection
  - Fraud Detection
  - Spam Detection
  - Data Mining for Homeland Security
- Adversary adapts to avoid being detected.
  - Millions different ways to write Viagra!
- New solutions are needed to address this problem

Understanding Adversarial Learning

- It is not concept drift
- It is noAdversary changes the distribution to avoid being detected
- t online learning
- There is game between the data miner and the adversary

Solution Ideas

- Constantly adapt your classifier to changing adversary behavior
  - Look at the Dalvi et.al. KDD 04 paper for such a solution for Naïve Bayes Classifier
- Questions??
  - How to model this game?
  - Does this game ever end?
  - Is there an equilibrium point in the game?

Adversarial Stackelberg Game

- Usually classifier is modified after observing adversaries action.
  - Spam filter rules.
  - Searches at metro stations at NY city.
- Stackelberg Games
  - Adversary chooses an action a_1
  - After observing a_1 , data miner chooses action a_2
  - Game ends with payoffs to each player

Our Formulation

- Two class problem
  - Good class, Bad class
- Mixture model

$$x = (x_1, x_2, x_3, \ldots, x_n)$$

$$p_1 + p_2 = 1$$

$$f(x) = p_1 f_1(x) + p_2 f_2(x)$$

- Adversary applies a transformation T to modify bad class $f_2(x) \xrightarrow{T} f_2^T(x)$
- After observing transformation, data miner chooses an updated classifier h
- We define the payoff function for the data miner

$$f(x) = p_1 f_1(x) + p_2 f_2^T(x)$$

$$c(T,h) = \int_{L_1^h} c_{11} p_1 f_1(x) + c_{12} p_2 f_2^T(x) dx + \int_{L_2^h} c_{21} p_1 f_1(x) + c_{22} p_2 f_2^T(x) dx$$

$$u_2(T,h) = -c(T,h)$$

- C_{ij} is the cost for classifying x to class i to given that it is in class j
- Data miner tries to minimize c(T,h)
- Transformation has a cost for the adversary
  - Reduced effectiveness for spam e-mails
- Let $g^T(x)$ be the gain of an element after transformation
- Adversary gains for the "bad" instances that are classified as "good"

$$u_1(T,h) = \int_{L_1^h} g^T(x) f_2^T(x) dx$$

- Given the transformation T, we can find the best response classifier(R(T)) h that minimizes the c(T,h)

$$h_T(x) = \begin{cases} \pi_1, (c_{12} - c_{22}) p_2 f_2^T(x) \le (c_{21} - c_{11}) p_1 f_1(x) \\ \pi_2, \text{otherwise} \end{cases}$$

- For Adversarial Stackelberg game, subgame perfect equilibrium is:

$$T^* = \arg\max_{T \in S} (u_1(T, R(T)))$$

$$(T^*, R(T^*))$$

UTD

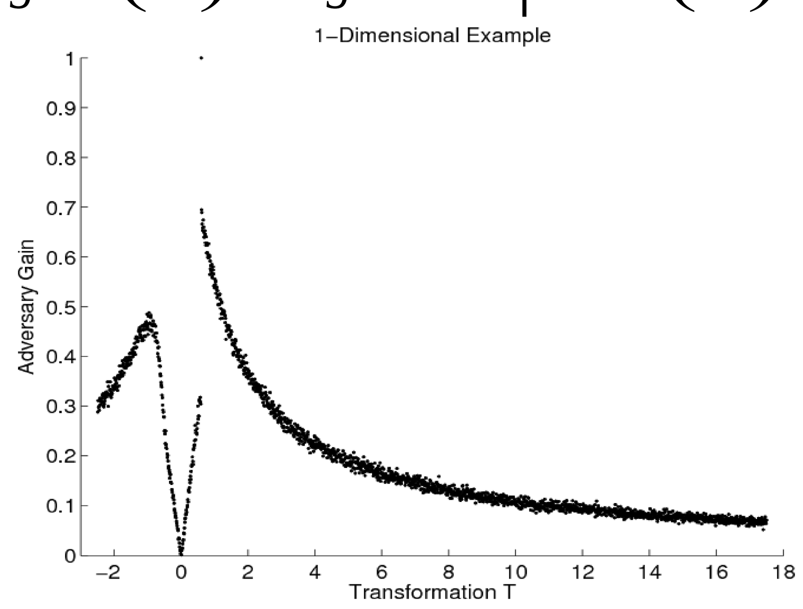
Solving For Equilibrium

- It is even hard to calculate $g_e(T)$ for given T
- Hard to maximize the $g_e(T)$
- Stochastic Optimization Ideas:
  - Monte-Carlo Integration
  - Simulated Annealing
- After an equilibrium is reached, each party does not have incentive change their actions.

Simulations for Mixture Models

- T is the set of all linear transformations
- Each class is assumed to be the Gaussian distribution.
- Cost of transformation for the adversary is

$$g^T(x) = g - a\left|T^{-1}(x) - x\right|_1$$



Attribute Selection for Adversarial Learning

- How to choose attributes for Adversarial Learning?
  - Choose the most predictive attribute
  - Choose the attribute that is hardest to change

Att.	f1()	f2()	Penalty	Equlibrium Bayes Error
X1	N(1,1)	N(3,1)	a=1	0.16
X2	N(1,1)	N(3.5,1)	a=0.45	0.13
X3	N(1,1)	N(4,1)	a=0	0.23

- Choose the attribute with best equilibrium performance!!

PURDUE UNIVERSITY

CERIAS

Discovery Park e-Enterprise Center