A Game Theoretical Model for Adversarial Learning

Author: Wei Liu, weiliu@it.usyd.edu.au
Supervisor: Prof. Sanjay Chawla
School of Information Technologies

If you know your enemies and know yourself, you can win a hundred battles without a single loss.

-- Sun Tzu, *The Art of War*, 500 BC

1. Motivation

In many situations where classifiers are deployed, adversaries *deliberately manipulate* data in order to reduce the classifier's accuracy. The most prominent example is email spam, where spammers routinely modify emails to get past classifier-based spam filters.

After observing the adversaries' strategies, extensive investigations that bring forward *anti-manipulation* classifiers such as robust spam filters are in high demand.

2. Methods

We introduce a new *Stackelberg game* to model the interaction between the adversary and the data miner, and show how to infer the equilibrium strategy.

We propose the use of *genetic algorithms* to solve the Stackelberg game where the players do not know each other's payoff function.

3. Definitions

- A game is played between two players: the Leader *L* (the spammer) and the Follower *F* (the data miner).
- Define sets of actions (strategies), *U* and *V* for *L* and *F* respectively.
- We denote payoff functions by J_L and J_F such that each J_i (i=L,F) is a twice-differentiable mapping $J_i(U, V) \rightarrow R$, where R stands for reaction.
- Thus the reaction function of *L* is:

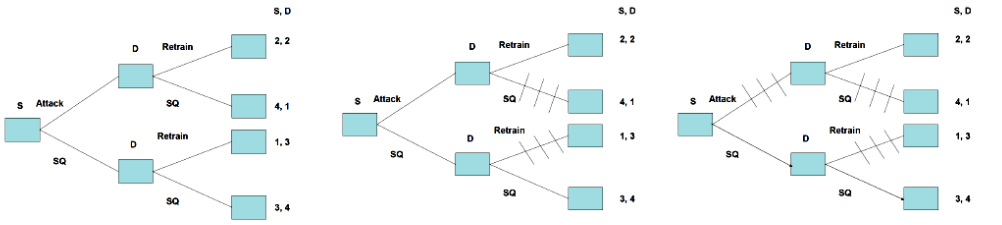
$$R_L = \arg\max_{v \in V} J_L(u, v)$$

4. Classification Problems

- We denote spams by $P(\mu_1, \sigma)$ and legitimate emails by $Q(\mu_2, \sigma)$. Then *L* plays by moving μ_1 to μ_1+u (towards μ_2) as shown in Figure 2b.
- We use Kullback-Leibler divergence to estimate the effects of *u* on *P*:

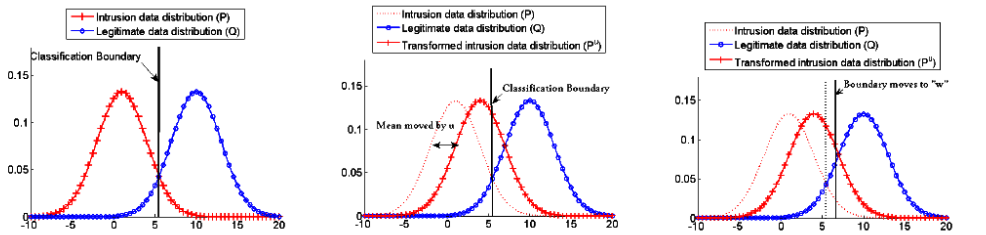
$$D_{KL}(N_1|N_2) = \frac{1}{2}(\log_e(\frac{det\Sigma_2}{det\Sigma_1})) + tr(\Sigma_2^{-1}\Sigma_1) + (\mu_2 - \mu_1)^{Transpose}\Sigma_2^{-1}(\mu_2 - \mu_1) - q$$

where *q* is the number of features in an attribute.



(a) Four outcomes with payoff sets from four different combinations of strategies.
(b) Two leaves are pruned from the perspective of data miner's payoff.
(c) The upper branch is pruned by the spammer based on the data miner's possible reactions.

Figure 1: Game tree for Stackelberg model between the spammer (*S*) and the data miner (*D*). "SQ" stands for status quo; "Retrain" means retraining the classifier.



(a) Initial model state
(b) The adversary's movement
(c) The data miner's reaction to the adversary's movement

Figure 2: Three status of the game theoretical model in classification scenario. The vertical lines represent the classification boundary built by naive bayesian classifier.

5. Genetic Algorithms (GA)

- We use GA to solve for the reaction function of *L*, where the solution yields the Stackelberg equilibrium game.
- The final generation of GA contains the best transformations for an adversary.

7. Conclusions

- The data miner should make more use of features that are more expensive to be transformed.
- The adversary will tend to make further transformation at Stackelberg equilibrium when the penalty of reconstructing classifiers is high.

		u	w	J^L	J^F	ErrRate	FPR	FNR
$\alpha = 0$	$\beta = 0$	9	10	0.5	0	50.00%	50.00%	50.00%
	$\beta = 0.01$	9	5.5017	0.9877	0	50.00%	1.22%	98.77%
	$\beta = 0.1$	9	5.5	0.9878	0	50.00%	1.22%	98.78%
$\alpha = 0.01$	$\beta = 0$	0.8229	5.9114	0.0148	1.9181	2.05%	2.05%	2.05%
	$\beta = 0.01$	1.1598	5.9937	0.0175	1.8972	2.51%	2.26%	2.76%
	$\beta = 0.1$	7.8368	5.9463	0.4652	0.0858	47.36%	2.13%	92.58%
$\alpha = 0.05$	$\beta = 0$	0.3943	5.6971	0.0138	1.9371	1.57%	1.57%	1.57%
	$\beta = 0.01$	0.5069	5.7069	0.0147	1.9320	1.69%	1.59%	1.79%
	$\beta = 0.1$	1.6520	5.8346	0.0217	1.8399	3.72%	1.86%	5.58%
$\alpha = 0.1$	$\beta = 0$	0.1749	5.5875	0.0129	1.9453	1.37%	1.37%	1.37%
	$\beta = 0.01$	0.2179	5.5869	0.0133	1.9437	1.41%	1.37%	1.45%
	$\beta = 0.1$	0.3752	5.5556	0.0148	1.9368	1.57%	1.31%	1.83%

Table 2: Variations of Stackelberg Equilibrium with different combinations of α and β when $\mu_1 = 1$, $\mu_2 = 10$ and $\sigma = 2$

6. Experiments

We use the parameter α to determines the strength of the KLD penalty, and β to control the strength of the cost of the classifier's boundary adjustment. The effects of these two parameters are shown in the above table.

8. References

1) Wei Liu and Sanjay Chawla, A Game Theoretical Model for Adversarial Learning. To appear in *Proceedings of the 2009 IEEE International Conference on Data Mining Workshops* (***ICDMW'09***), Miami, FL, USA, December 6 – 9, 2009.
2) Sun Tzu. The Art of the War, 500bc.

This research is sponsored by Capital Market CRC, and Australian Research Council

The University of Sydney