

STANFORD UNIVERSITY
DEPARTMENT OF STATISTICS
DEPARTMENTAL SEMINAR

4:15 p.m., Tuesday, July 25, 2006
Sequoia Hall Room 200
(Cookies at 3:45 in 1st Floor Lounge)

Bowei Xi
Purdue University

Adversarial Learning

Many data mining applications, ranging from spam filtering to intrusion detection, are faced with active adversaries. In all these applications, initially successful classifiers could degrade easily. This becomes a game between the adversary and the data miner: The adversary modifies its strategy to avoid being detected by the current classifier; the data miner then updates its classifier based on the new threats. We investigate the possibility of an equilibrium in this seemingly never ending game, where neither party has an incentive to change: Modifying the classifier causes too many false positives with too little increase in true positives; changes by the adversary decrease the utility of the false negative items that aren't detected.

We develop a game theoretic framework where the equilibrium behavior of adversarial learning applications can be analyzed, and provide a solution for estimating the equilibrium point. The data miner could evaluate the eventual effectiveness of its current approach, and decide whether it is necessary to change the rules of the game (e.g., find new information as input to the classification problem). We also describe how to apply our techniques in a spam filtering scenario.

This is joint work with Murat Kantarcioglu and Chris Clifton.