

CS905: Project Proposal

20th Jan, 2007

Adversarial Classification Problem

Submitted By:-

Deepti Bhatnagar (2002433)
Gaurav Gupta (2002435)
Varun Nayyar (2002445)

Background

Essentially all data mining algorithms assume that the data-generating process is independent of the data miner's activities. However, in many domains, including spam detection, intrusion detection, fraud detection, surveillance and counter terrorism, this is far from the case. The data is actively manipulated by an adversary seeking to make the classifier produce false negatives. In these domains, the performance of a classifier can degrade rapidly after it is deployed, as the adversary learns to defeat it.

Present Scenario

A brief literature survey has yielded that currently, the only solution to this problem is a repeated, manual, ad hoc reconstruction of the classifier. Many researchers are still open to more promising approaches.

Our Approach

We plan to look upon classification as a game between the classifier and the adversary, and produce a classifier that is optimal given the adversary's optimal strategy. Experiments in a spam detection domain have shown that this approach can greatly outperform a classifier learned in the standard way, and (within the parameters of the problem) automatically adapt the classifier to the adversary's evolving manipulations. Our methodology has been largely influenced by the *Nilesh Dalvi, Pedro Domingos, Mausam, Sumit Sanghai, Deepak Verma's* Adversarial classification. Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD), 2004: 99-108.

Resources

From the likely initial estimates of this category of implementation we will not need any additional resources like hard disks, additional CPU's. The data for the same should be manageable and not pose any additional burden on the system administrators.