DEFENDING ACTIVE LEARNING AGAINST MALICIOUS INPUTS IN AUTOMATED DOCUMENT CLASSIFICATION

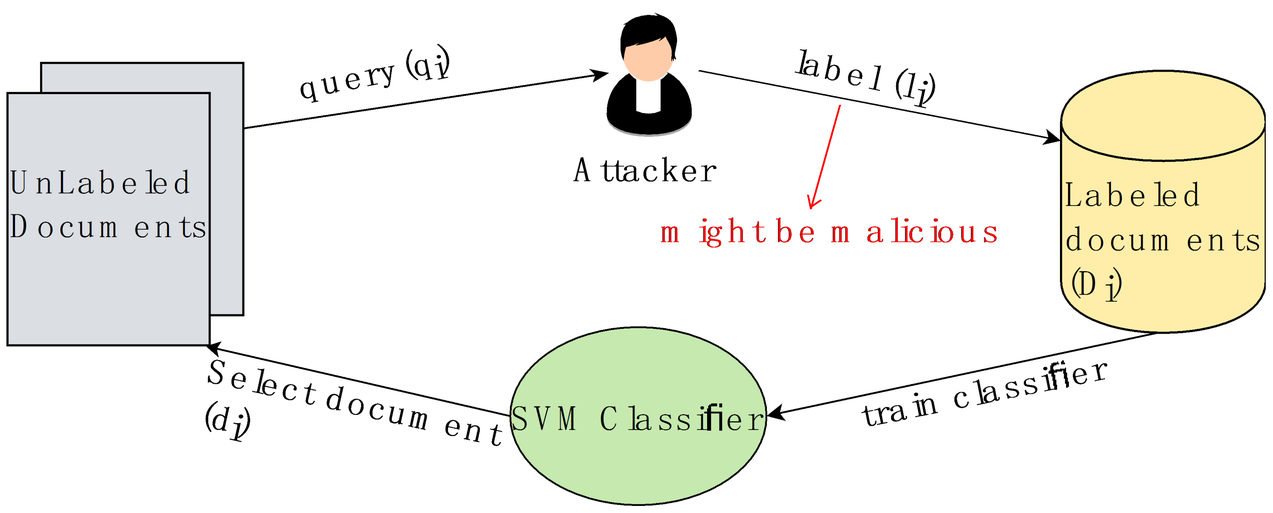
Lei Pi — *University of Memphis*
Zhuo Lu — *University of South Florida*
Yalin Sagduyu — *Intelligent Automation Inc.*
Su Chen — *University of Memphis*

Objective:

Accuracy of a classifier is important in large-volume automated document categorization but subject to malicious human inputs. This paper aims to design a method to identify erroneous labeled data and alleviate the damage to classifiers caused by these malicious inputs.

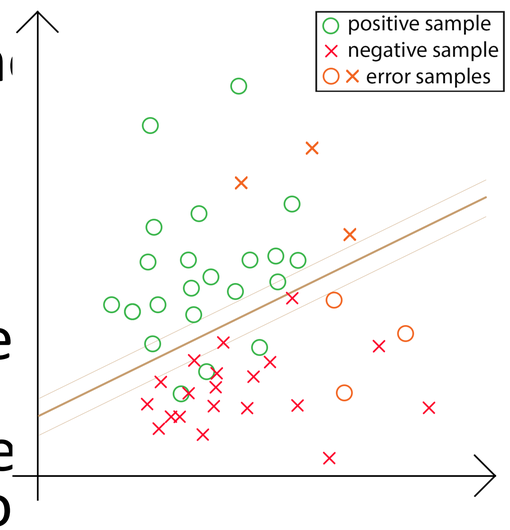
Scenario:

- In active learning, samples are selected from unlabeled document set and passed to oracles
- An oracles labels samples and adds result to labeled document set as training set
- An SVM classifier is trained with the labeled documents
- An attacker provides falsified labels to misguide the training process.



Intuition:

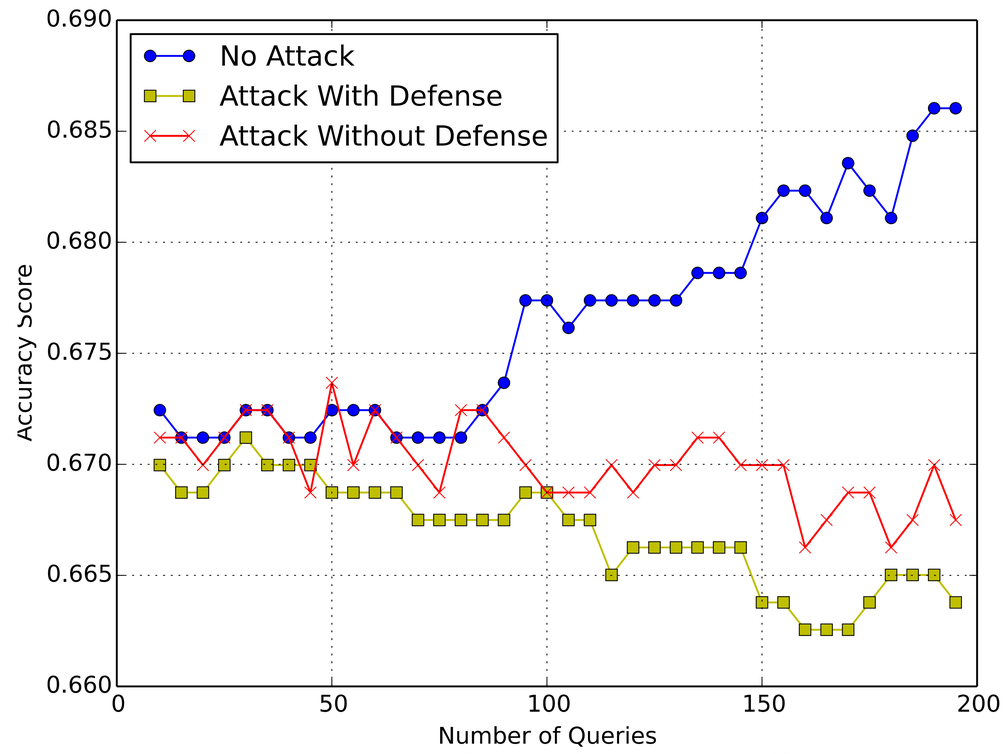
- Attacker does not know the position of the separating plane.
- If a mislabeled sample is too far away from the separating plane, it's more likely to be malicious.
- Find a risk factor *r* to judge how far a label could be to the hyperplane to remain legit.



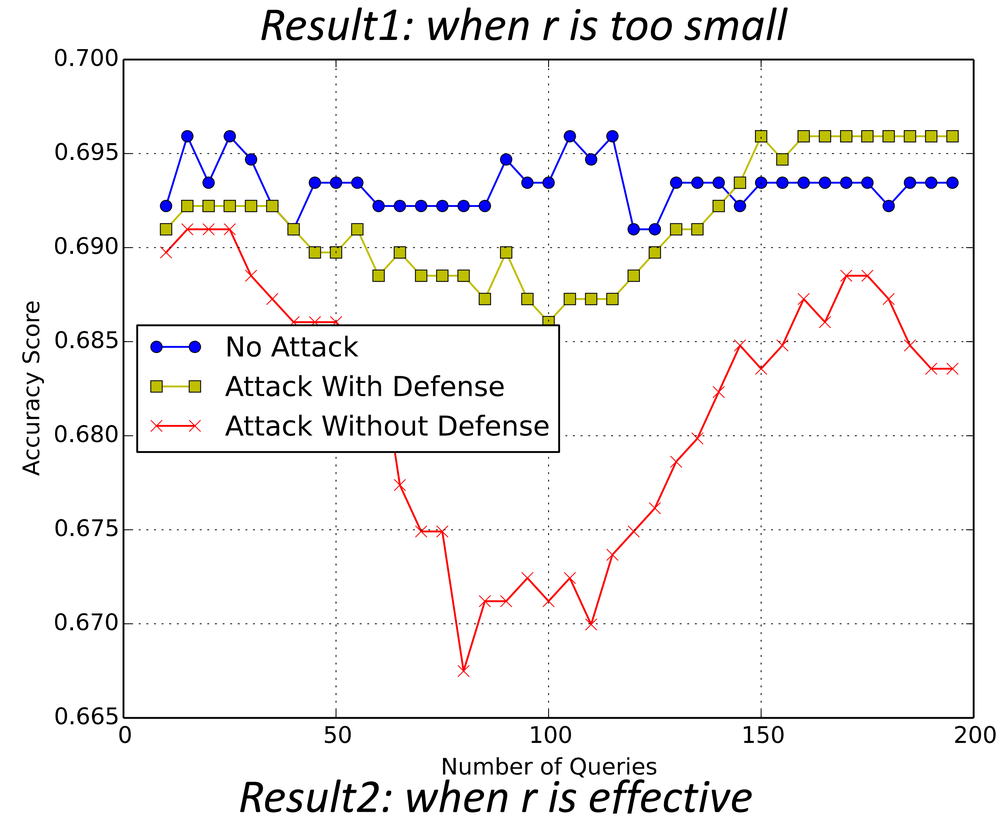
Setup:

- *Dataset*: 1264 instances with 10233 extracted features from documents in Reuters-21578.
- *SVM kernel*: RBF with λ=1.0/1264 and C = 1.0
- *Erroneous labeling rate*: 25%
- *Defense settings*: no attack, attack with defense, attack without defense

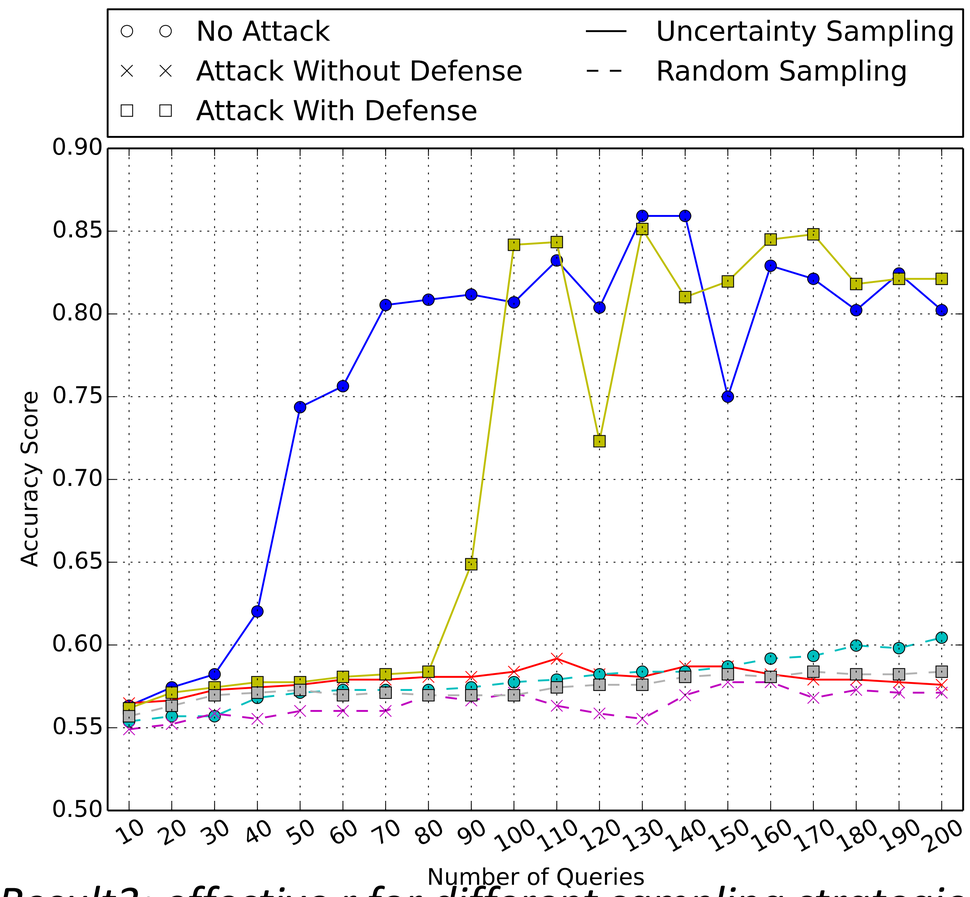
Results:



Result1: when r is too small



Result2: when r is effective



Result3: effective r for different sampling strategies

Future works:

- Attacker identification and prevention
- More scalable algorithms to find *r*