

Ensembles in Adversarial Classification for Spam

Deepak Chinavle, Pranam Kolari, Tim Oates and Tim Finin

University of Maryland, Baltimore County

The Problem with Spam

- Web and email spam is a constant problem
- Effective spam classifiers can be built with sufficient, labeled training data
- As spammers change their tactics, however, the classifiers must be retrained
- Knowing when to retrain is a problem and obtaining new labeled data is expensive

Ensemble of Classifiers Approach

- Use an ensemble with one classifier per feature set extracted from data
- Changes in mutual agreement between labels assigned by pairs of individual classifiers in the ensemble indicate concept drift
- Retrain drifting classifiers using ensemble labels, no need for hand-labeled data

Evaluation

- Evaluated approach using spam blog data collected in 2005 and 2006 and hand labeled
- Used features from Kolari 2006
- Compare automatic retraining and using true labels
- Measure accuracy and cost to retrain under several policies

Results

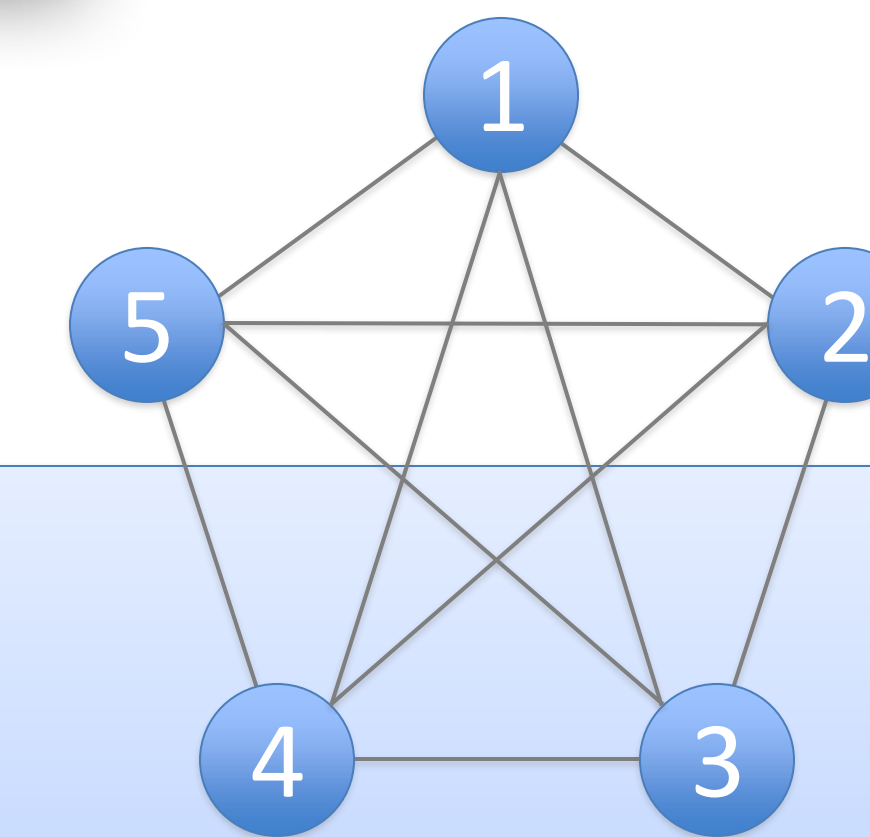
- Mutual agreement between classifiers is an accurate indicator of the need to retrain
- Concept drift typically affects a subset of the features, so ensemble is robust
- Ensemble labels are almost as good as true labels for retraining individual classifier



An example of a spam blog on wordpress.com and software used to create spam blogs

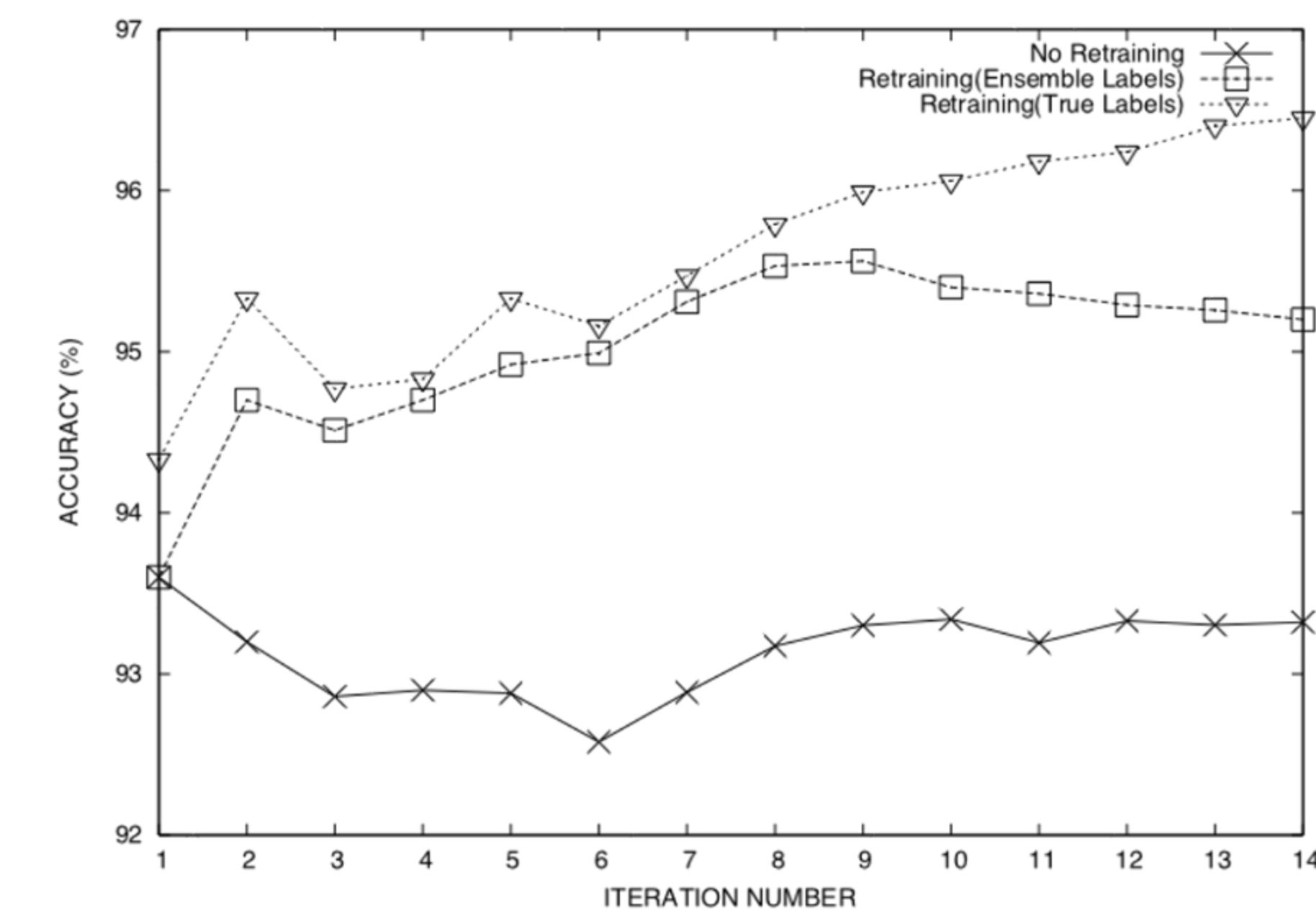
Mutual Agreement

- Percent of time a pair of classifiers agrees on label
- Evaluated for all pairs
- If one classifier is affected by concept drift but others are not, mutual agreement with that classifier will change
- No need for true labels to detect drift

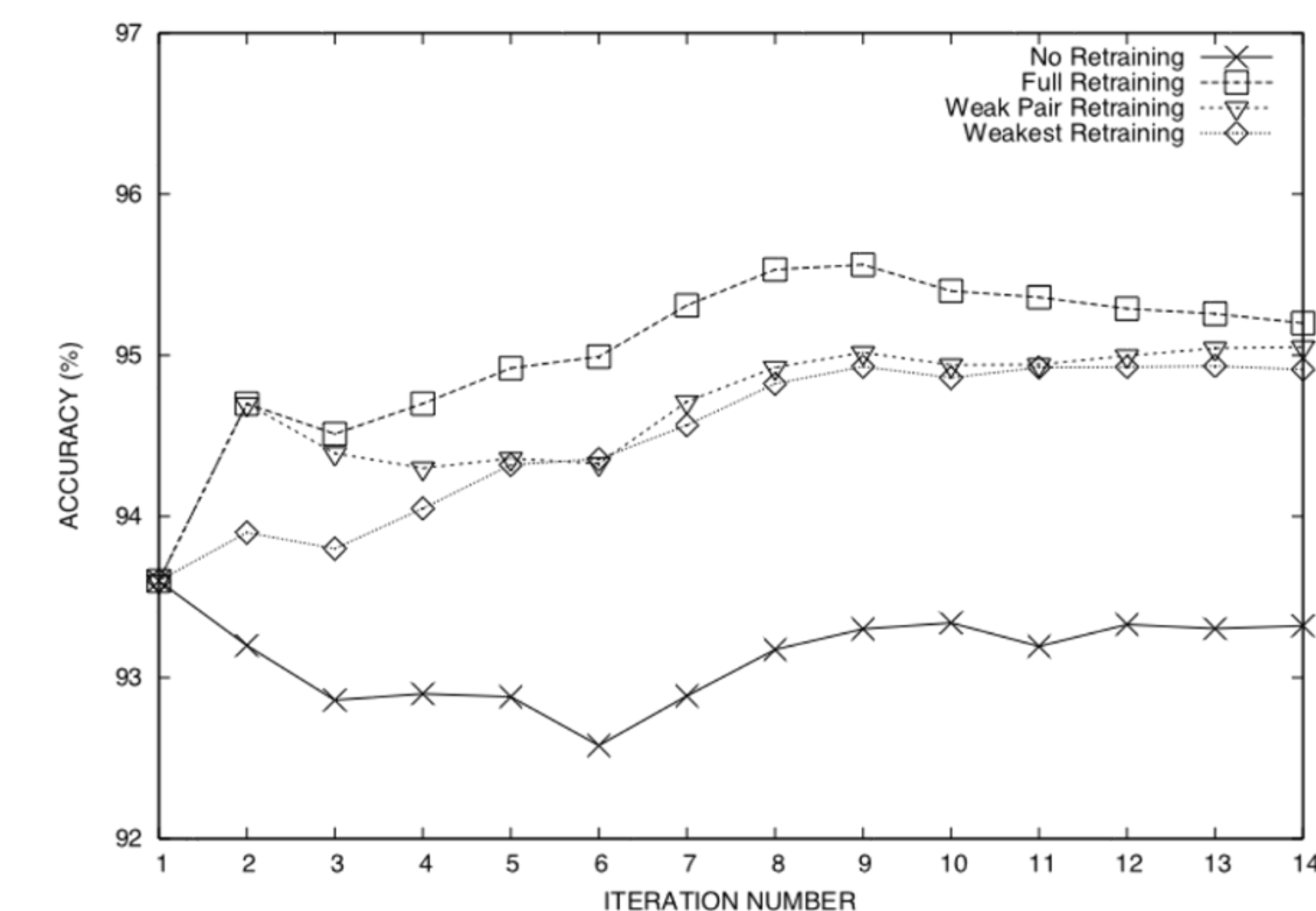


Example Feature Sets

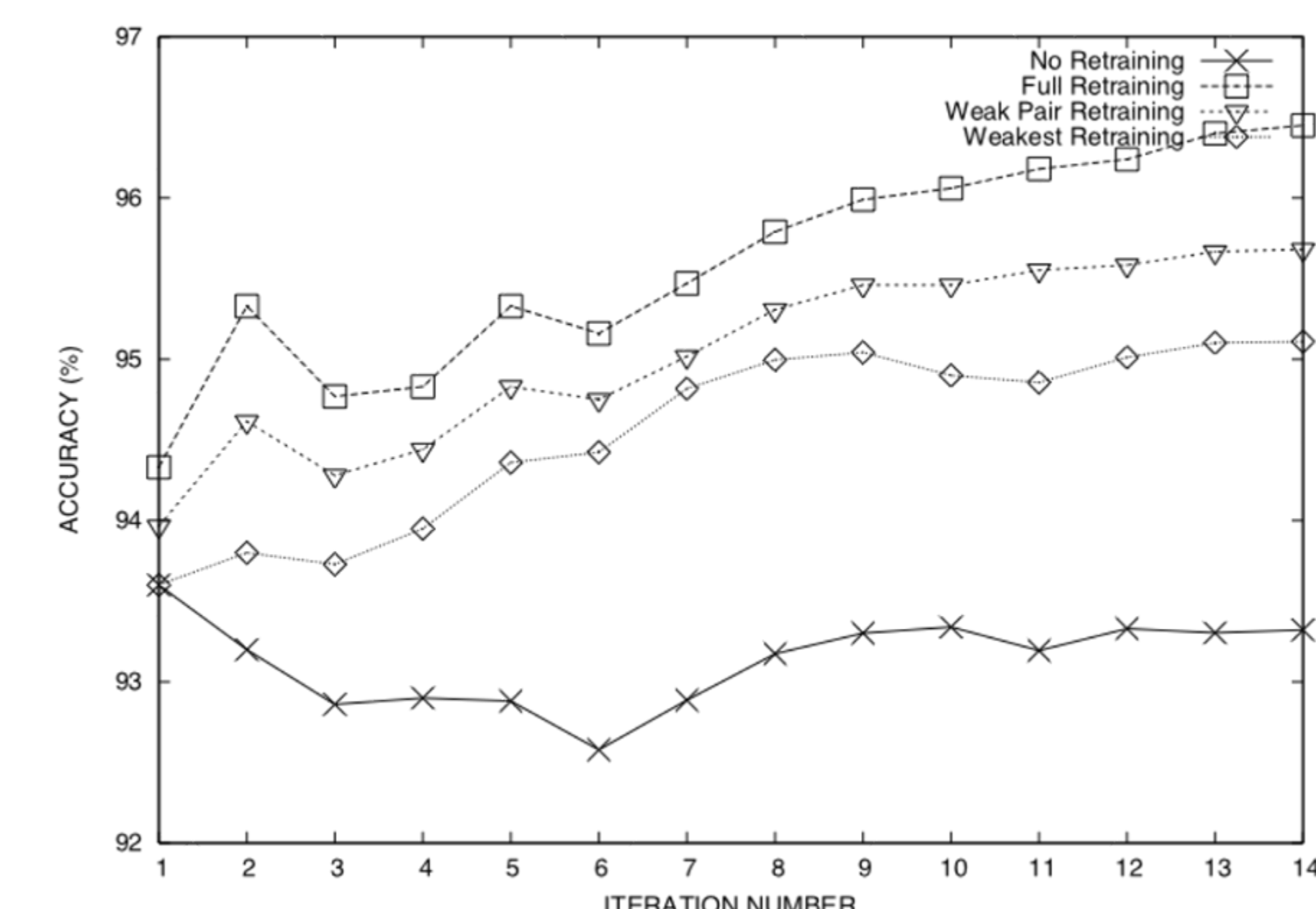
- Bag of words: word frequency in blog
- Word n-grams: frequency of short word phrases
- Character n-grams: frequency of short character sequences
- Anchor text: text in HTML links
- Tokenized URLs & outlinks: tokens extracted from links
- HTML tags: frequency of common tags, e.g., H1 and BOLD
- URL: tokens extracted from link to blog



Full retraining accuracy, ensemble vs. true labels and no retraining



Accuracy, ensemble labels for retraining



Accuracy, true labels for retraining