A Stopping Criterion for Active Learning

VLACHOS, A.
2008
A Stopping Criterion

- Manual annotation takes time and human effort
- Could stop when some performance is achieved
- A better solution would consider how much can be learnt by labelling unlabelled instances
- Proposed approach examines classifier confidence
Support Vector Machines (SVM)

- Typically used as a binary classifier
- Kernel functions (such as linear, $K(x_i, x_j) = x_i \cdot x_j$)
  - Compare instances
  - Effectively map to higher dimensions
- Classify by finding hyperplane with maximal margin
Training a Reuters SVM Classifier with AL

- SVMs trained on most popular topic in Reuters
- Margin used as measure of uncertainty
- Average margin of test set data used as a measure of the confidence of the classifier
- Three SVMs compared:
  - Random test data sampling
  - AL, adding 1% of pool to the training data each time
  - AL, adding 0.1% of pool to the training data each time
Linear SVM Training on Reuters

![Graphs showing F-score and average margin vs. % of training data used.](image)
The confidence follows a rise-peak-fall pattern

- Rise to a peak as training data with novel information is used (performance changes little after this)
- Falls as contradictory instances selected:
  - The classifier is confident, but incorrect, about these
  - Presumably, these are due to limitations of the feature set (e.g., a bag-of-words model ignoring word order)
Gaussian SVM Training on Reuters
Contradictory information is added when
- An instance whose label is incorrectly predicted
- Is added to the training data
- For round $t$ this is computed as:

$$Contradictory\_information(t) = \sum_{i \in i^t} \left| \frac{f^t(x_i)}{f^t(x)} \right|$$
SVM Contradictory Information
Linear SVM for an Infrequent Class

Graphs showing F-score and Average margin as a function of the % of training data used.
Experiments With Other Classifiers

Bayesian Logistic Regression Classifier

Maximum Entropy Classifier
Binary NER Classification
Set-up

- The dataset has four entity types to be recognised
- Reduce this to a binary task by just identifying if there is a named entity or not
- Train a linear kernel SVM using AL
  - Randomly choose 1% of data as seed data
  - Use 1% and 0.1% batches
- Classifier uses simple lexical features
Linear SVM for NER
Tricking the Stopping Criterion

- The criterion detects a rise-peak-fall pattern
- The criterion can be satisfied non-optimally with:
  - A noisy or misleading seed
  - Bad early selections
- Instead, require a consistent drop in the confidence
Multiclass SVM
The One-Against-All Scheme

This is a way to adapt SVMs for multiple classes

Approach

- Each classifier classifies true/false for one class
- Select the class which gives the largest positive margin

Define Confidence as the difference in the size of

- The most positive margin
- The next most positive margin
Set-up

- **NER experiments**
  - As before, but use the four separate classes

- **Shallow parsing experiments**
  - Goal is to divide text into chunks ("syntactically-related, non-overlapping groups of tokens")
  - Each token belongs to one syntactic category
  - 23 classes (with widely-varying numbers of instances)
  - Uses a previously-defined feature set
Multiclass SVM Experiments

Multiclass NER Classifier

Shallow Parsing Classifier
Conclusion
These experiments show how a stopping criterion based on the rise-peak-drop pattern could work.

We saw how this pattern appeared consistently with a variety of problems and classifiers.
Thorough explanation of background and context

Appears to be a novel, sensible, and effective extension of the then state-of-the-art

Should be particularly useful for NLP tasks

Formulation of the criterion wasn’t made very explicit

Comparisons to alternatives might be interesting