On feature distributional clustering for text categorization

Bekkerman, El-Yaniv, Tishby and Winter

Technion – Israel Institute of Technology

SIGIR 2001

Plan

- A new text categorization technique based on two known ingredients:
  - Distributional Clustering
  - Support Vector Machine (SVM)
- Comparative evaluation of the new technique with other works:
  - SVM + Mutual Information (MI) feature selection [Dumais et. al.]
  - SVM without feature selection [Joachims]

Main results

- The evaluation is performed on two benchmark corpora:
  - Reuters
  - 20 Newsgroups (20NG)
- The new technique outperforms others on 20NG.
- It does worse on Reuters.
- Possible reasons for this phenomenon.

Text categorization

- Supervised learning.
  - Categories are predefined.
- Many real-world applications.
  - Search engines.
  - Helpdesks.
  - E-mail filtering
  - More…

Text representation

- A standard scheme: Bag-Of-Words (BOW).
  - A document as a vector of word occurrences.
- A more sophisticated method: distributional clusters.
  - A word is represented as a distribution over the categories [McCallum, Pereira & Tishby & Lee].
  - The words are then clustered.
  - A document as a vector of centroid occurences.

Support Vector Machines

- A modern inductive learning scheme.
- Proposed by Vapnik.
- Usually shows advantage over other learning schemes such as
  - Naïve Bayes
  - *K*-Nearest Neighbors
  - Decision trees
  - Boosting

Corpora

- We've tested our algorithms on two well-known corpora:
  - **Reuters** (ModApte split): 7063 articles in the training set, 2742 articles in the test set. 118 categories.
  - **20 Newsgroups (20NG)**: 19997 articles. 20 categories.

Multi-labeling vs. uni-labeling

- Multi-labeled corpus: articles can belong to a number of categories.
  - Example: **Reuters** (15.5% are multi-labeled documents).
- Uni-labeled corpus: each article belongs to only one category.
- **20NG** has been often treated as uni-labeled. In fact it contains 4.5% multi-labeled documents.

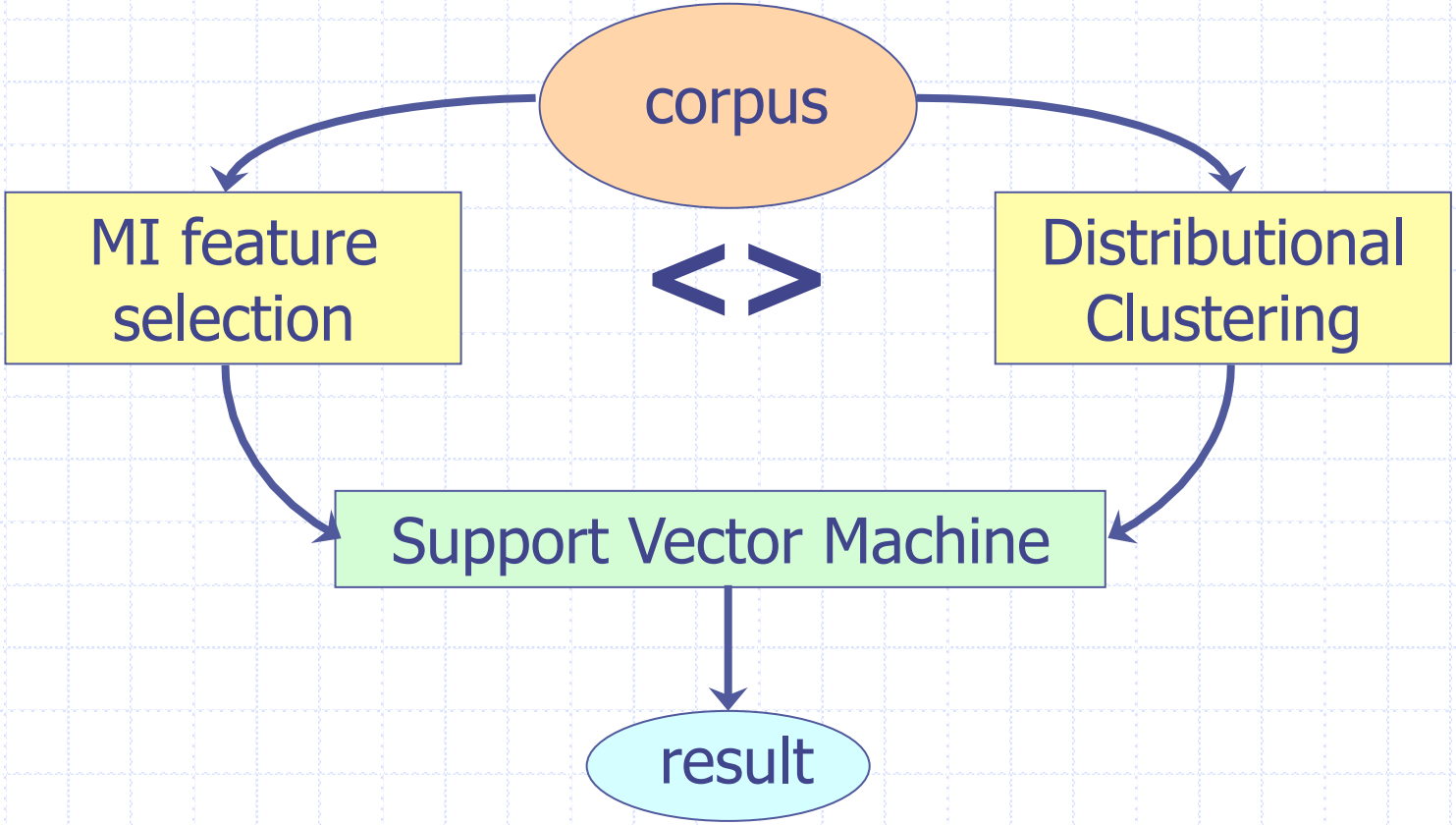
Some text categorization results

- Dumais et al. (1998): Linear SVM with simple feature selection on Reuters.
  - Achieve best known result: 92.0% of breakeven over 10 largest categories (multi-labeled).
- Baker and McCallum (1998): Distributional clustering + Naïve Bayes on 20NG.
  - 85.7% of accuracy (uni-labeled).

Results (contd.)

- Joachims (1996): Rocchio algorithm.
  - Best known result on 20NG (uni-labeled approach): 90.3% of accuracy.
- Slonim and Tishby (2000): Naïve Bayes + distributional clustering with small training sets.
  - Up to 18% of accuracy improvement over BOW on 20NG.

Our study


corpus
MI feature selection
<>
Distributional Clustering
Support Vector Machine
result

Feature selection via Mutual Information

- In training set, choose *k* words which best discriminate the categories.
- In terms of Mutual Information:

$$I(w,c) = \sum_{e_w \in \{0,1\}} \sum_{e_c \in \{0,1\}} p(e_w, e_c) \log \frac{p(e_w, e_c)}{p(e_w)p(e_c)}$$

  - For each word w and each category c

Feature selection via MI (contd.)

- For each category we build a list of *k* most discriminating terms.
- For example (on 20 Newsgroups):
  - ***sci.electronics:*** circuit, voltage, amp, ground, copy, battery, electronics, cooling, …
  - ***rec.autos:*** car, cars, engine, ford, dealer, mustang, oil, collision, autos, tires, toyota, …
- Greedy: does not account for correlations between terms.

Distributional Clustering

- Proposed by Pereira, Tishby and Lee (1993).
- Its generalization is called *Information Bottleneck (IB)* [Tishby, Pereira, Bialek 1999].
- In our case, each word (in the training set) is represented as a distribution over categories it appears in.
- Each word w is then clustered into a centroid ("pseudo-word") $\tilde{w}$.

Information Bottleneck (IB)

- The idea is to construct $\tilde{w}$ so that to maximize the Mutual Information $I(\tilde{w},c)$ under a constraint on $I(\tilde{w},w)$.
- The solution is in the following equation:

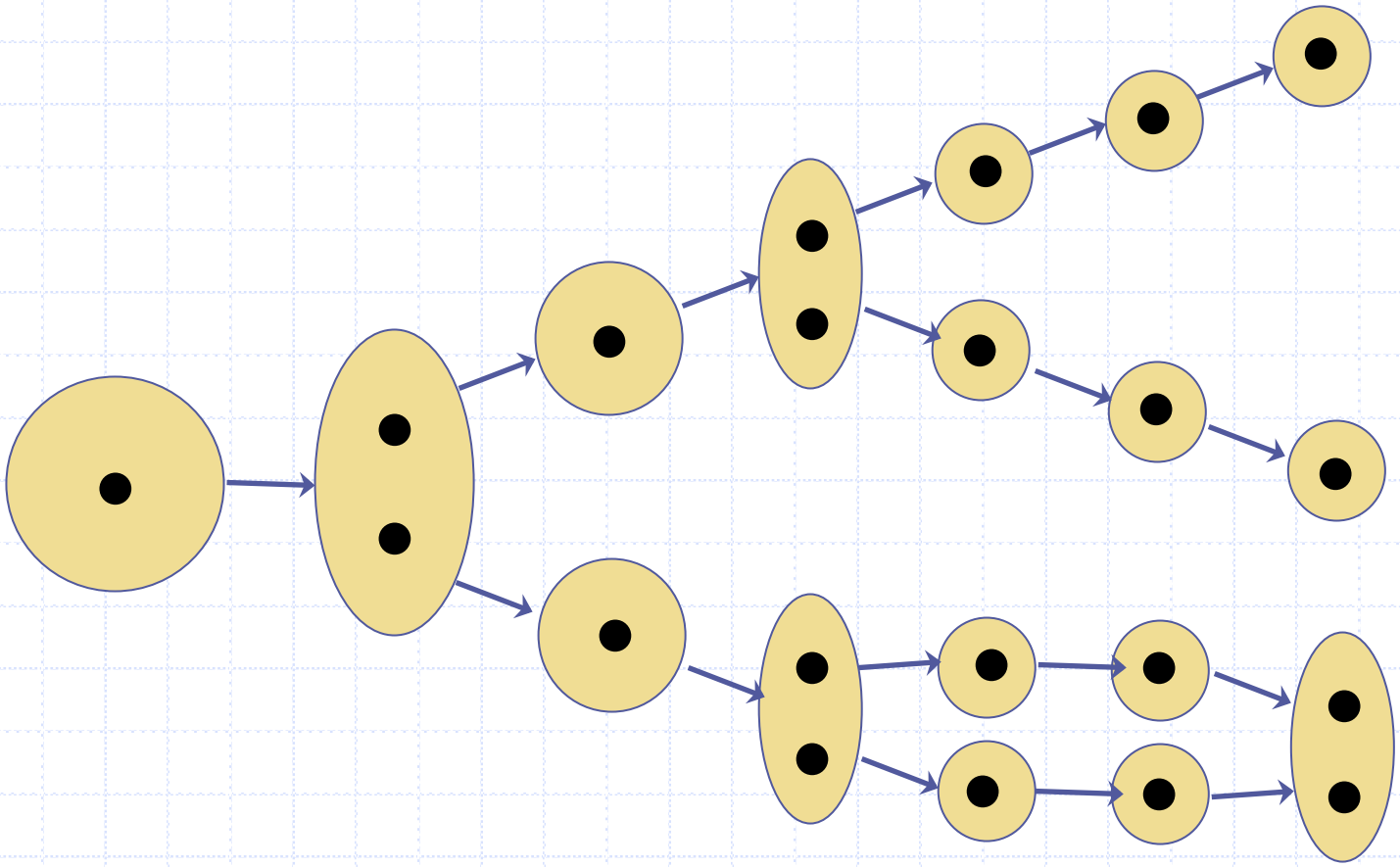
$$p(\tilde{w}\,|\,w) = \frac{p(\tilde{w})}{Z(\beta,w)} \exp\left[-\beta \sum_c p(c\,|\,w) \ln \frac{p(c\,|\,w)}{p(c\,|\,\tilde{w})}\right]$$

Z is the normalization factor, β is an annealing parameter.

Deterministic Annealing (DA)

- Solution for IB equations can be obtained using a clustering routine similar to DA.
- DA: a powerful clustering method, proposed by Rose et. al. (1998).
- The approach is "top-down":
  - Start with one cluster with low β ("high temperature").
  - Split it while lowering the "temperature" until reaching a stable stage.

Deterministic Annealing (contd.)

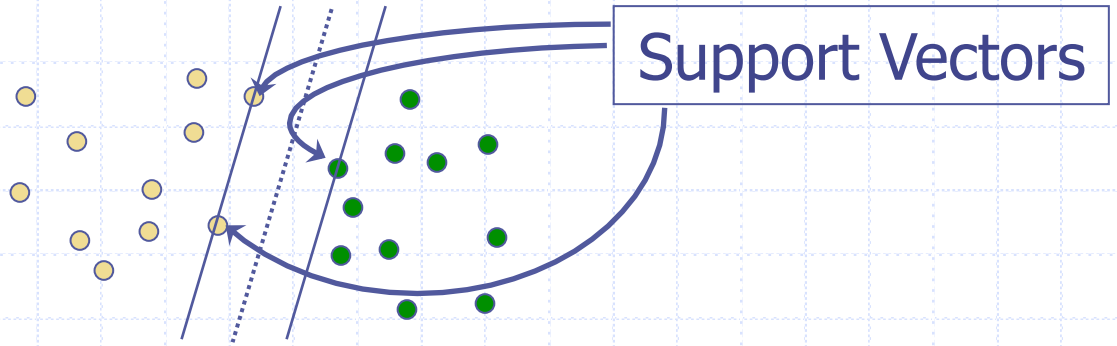


Document Representation

- In MI feature selection technique:
  - Documents are projected onto *k* most discriminating words.
- In Information Bottleneck technique:
  - At first words are grouped into clusters,
  - And then documents are projected onto the "pseudo-words".
- So, documents are vectors whose elements are numbers of occurrences of *"best"* words (1) or *"pseudo-words"* (2).

Support Vector Machines

- Goal: find a decision boundary with maximal margin.



- We used linear SVM (implementation: SVM*light* by Joachims).

Multi-labeled categorization via binary decomposition

1. MI feature selection (or distributional clustering) on the training and test sets.
2. For each category we train a binary classifier on the training set.
3. On each document in the test set we run all the classifiers.
4. The document is related to all the categories whose classifiers accepted it.

Uni-labeled categorization via binary decomposition

1. MI feature selection (or distributional clustering) on the training and test sets.
2. For each category we train a binary classifier on the training set.
3. On each document in the test set we run all the classifiers.
4. The document is related to the (one) category whose classifier accepted it with maximal score (*"max-win"* scheme)

the same as in multi-labeled scheme

Evaluating the results

- **Multi-labeled**: each document's labels should be identical to the classification results.
  - Precision/Recall/Breakeven/**F**-measure
- **Uni-labeled**: the classification result should match the true label, or be in the set of true labels.
  - Accuracy measure (number of "hits").

Experimental setup

- To reproduce the results achieved by Dumais et. al., we took $k = 300$ (number of "best" words and number of clusters).
- Since we wanted to compare 20NG and Reuters (ModApte split: ¾ is training set and ¼ is test set) we used **4**-fold cross-validation on 20NG.

Parameter tuning

- We have 2 major sets of parameters:
  - Number of clusters or "best" words (*k*).
  - SVM parameters (*C* and *J* in SVM*light*).
- For each experiment, *k* is fixed.
- To perform a "fair" experiment, we tune *C* and *J* on a validation set (splitting the training set into **train-train** and **train-validation** subsets).
- Then we run the experiment with the best parameters found.

Unfair parameter tuning

- Suppose we want to compare performance of two classifiers *A* and *B*.
- To empirically show that *A* is better than *B*, it is sufficient to
  - Tune *A*'s parameters as described above (validation set)
  - Tune *B*'s parameters in an unfair manner (over the test set)

Results on 20 Newsgroups

- Multi-labeled setting (breakeven point):
  - Clustering: 88.6±0.3% (k = 300)
  - MI feature selection: 78.9±0.5% (k = 300)
  - `` `` : 86.3±0.4% (k = 15000)
- Uni-labeled setting (accuracy measure):
  - Clustering: 91.2±0.6% (k=300)
  - MI feature selection: 85.1±0.5% (k = 300)
  - `` `` : 91.0±0.2% (k = 15000)
- Parameter tuning of the MI-based experiments is **unfair**.

Result on Reuters

- Multi-labeled setting (breakeven point):
  - Clustering: 91.2% (k = 300)
    - Unfair: 92.5%
  - MI feature selection: 92.0% (k = 300) as published by Dumais et al.
- The results are achieved on 10 largest categories of Reuters.

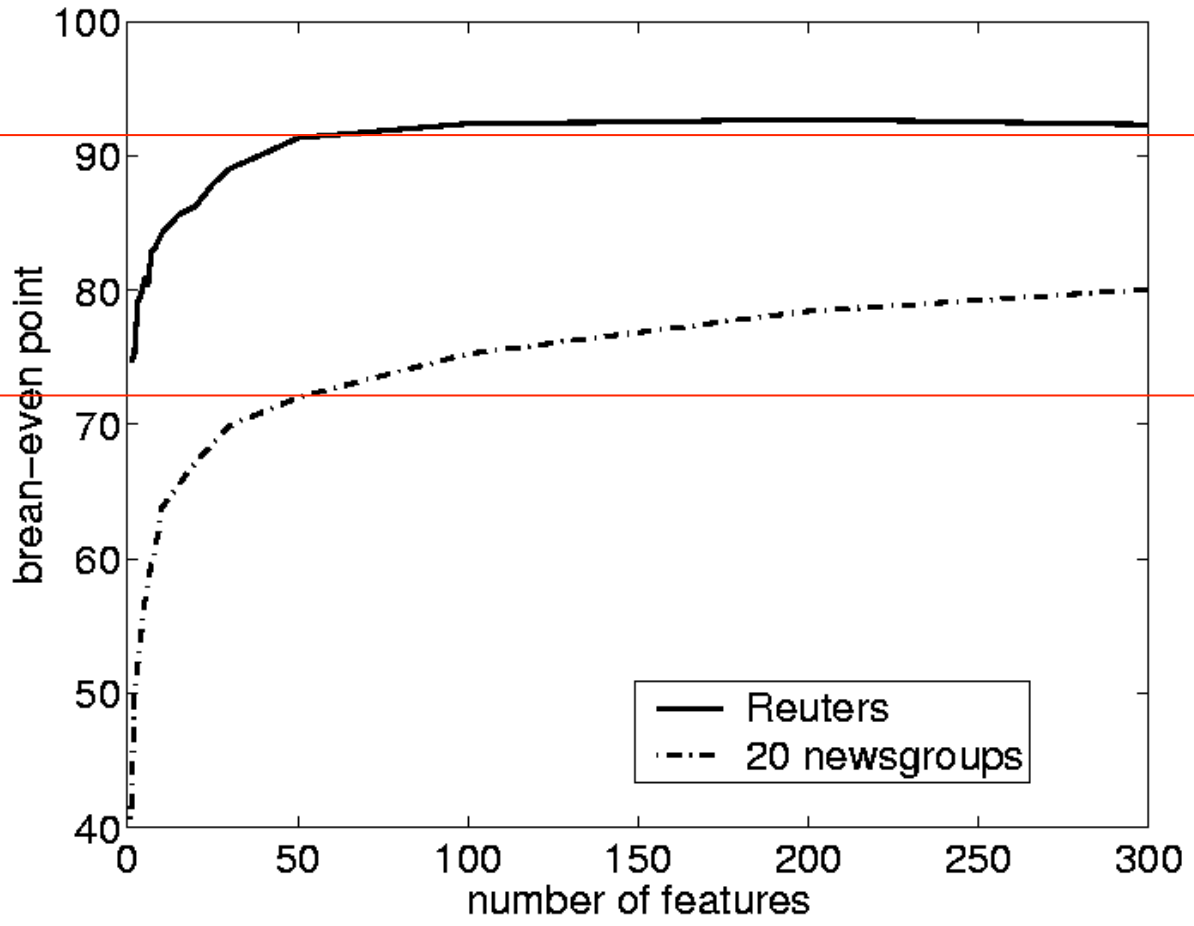
Discussion of the results

- On 20NG our technique (clustering) is either:
  - more accurate than MI
  - OR more efficient than MI
- On Reuters: a little worse. Why?
- *Hypothesis:* Reuters was labeled only according to a few **keywords** that appeared in the documents. 20NG articles wee labeled by their authors, based on full understanding

BEP vs. Feature set size

- We examined performance as a function of number of features.
- We saw that
  - On 20NG the results increased sharply,
  - On Reuters the results remained the same.
- So, just a few words are enough to categorize documents of Reuters, while in 20NG we need much more words.

Dependence of BEP on number of features



Bekkerman, El-Yaniv, Tishby, Winter – SIGIR 2001

Example: BEP on 3 features

Category	1st word	2nd word	3rd word	BEP
Earn	vs+	cts+	loss+	93.5%
Acq	shares+	vs-	Inc+	76.3%
Money-fx	dollar+	vs-	exchange+	53.8%
Grain	wheat+	tonnes+	grain+	77.8%
Crude	oil+	bpd+	OPEC+	73.2%
Trade	trade+	vs-	cts-	67.1%
Interest	rates+	rate+	vs-	57.0%
Ship	ships+	vs-	strike+	64.1%
Wheat	wheat+	tonnes+	WHEAT+	87.8%
Corn	corn+	tonnes+	vs-	70.3%

Concluding Remarks

- SVM+IB method on 20NG is
  - either more efficient
  - or more accurate
- For Reuters: BOW is the best!
  - Don't try your fancy representation methods
- Open: can one devise a "universal" representation method that is best on all corpora?