Feature Selection in Text Categorization

[historical view]
Y. Yang & J. Pedersen
ICML, 1997

Motivation and Goals

- Text categorization problems typically have high dimensional feature spaces
  - Would be good to reduce the feature set size without sacrificing categorization accuracy
- Perform a comparative study of feature selection methods for text categorization
  - Focus on aggressive dimensionality reduction
  - Examine 5 methods

Research questions

- What are the strengths and weaknesses of existing feature selection methods?
- To what extent can feature selection improve the accuracy of a classifier? How much can we reduce the vocabulary without losing useful information for category prediction?

Feature selection methods

- Each uses a term-goodness criterion
- Thresholded to achieve the desired degree of term elimination
Feature selection methods

- Document frequency thresholding (DF)
  - DF is the # of documents in which a term occurs
  - Remove from the feature space those terms with DF < threshold (predetermined)
  - Simplest of the techniques explored
  - Issue: in ad-hoc retrieval tasks, low-DF terms are assumed to be informative !!

Feature selection methods

- Information gain (IG)
  - The best features are those that discriminate among the various classes
  - Binary case: CS major database example

<table>
<thead>
<tr>
<th>Height</th>
<th>Eyes</th>
<th>Class</th>
</tr>
</thead>
<tbody>
<tr>
<td>short</td>
<td>brown</td>
<td>hacker</td>
</tr>
<tr>
<td>tall</td>
<td>blue</td>
<td>theoretician</td>
</tr>
<tr>
<td>short</td>
<td>blue</td>
<td>theoretician</td>
</tr>
</tbody>
</table>

Feature selection methods

- Mutual information (MI)
  - Used in NLP to model word associations
  - Examines the # of times two words co-occur vs. the # of times they occur independently
  - One problem with MI: favors rare terms

Feature selection methods

- Chi-squared statistic (CHI)
  - Measures the lack of independence between a term and a category
  - Not reliable for low-frequency terms
### Feature selection methods

- **Term strength**
  - Estimates term importance based on how commonly a term is likely to appear in “closely-related” documents.
  - Quite different from the other methods.
  - Based on document clustering: documents with many shared words are related; terms shared between related documents are relatively important.

### Classifiers

- **kNN:** k-nearest-neighbor
  - Weighted
- **LLSF:** linear least squares fit regression
- Both were considered good methods at the time.

### Data

- **Reuters-22173**
  - 9610 training; 3662 testing
  - 92 categories
  - 1.24 categories per document
  - 16,039 terms

### OHSUMED

- Subset of MEDLINE
  - 14,321 categories
  - 1990 abstracts: training
    - 72,076 terms
  - 1991 abstracts: testing
    - Average of 12 categories per document

### Evaluation

- Recall
- Precision
- 11 point average precision
Term weighting

Performance curve: k-NN

<table>
<thead>
<tr>
<th>Term Frequency</th>
<th>Inverse Document Frequency</th>
<th>Normalization</th>
</tr>
</thead>
<tbody>
<tr>
<td>First Letter</td>
<td>$f(tf)$</td>
<td></td>
</tr>
<tr>
<td>Second Letter</td>
<td>$f(df)$</td>
<td></td>
</tr>
<tr>
<td>Third Letter</td>
<td>$f(length)$</td>
<td></td>
</tr>
<tr>
<td>n (natural)</td>
<td>$t(f)$</td>
<td></td>
</tr>
<tr>
<td>(logarithmic)</td>
<td>$1 + \log(tf)$</td>
<td></td>
</tr>
<tr>
<td>a (augmented)</td>
<td>$0.5 + 0.5 \times \frac{f}{\max_{i=1}^{n}}$</td>
<td></td>
</tr>
</tbody>
</table>

Table 1: Term Weights in the Smart System

Performance curve: LLSF

Qualitative comparison

<table>
<thead>
<tr>
<th>Method</th>
<th>DT</th>
<th>LG</th>
<th>CPH</th>
<th>M</th>
<th>TS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Inversion common terms</td>
<td>Y</td>
<td>Y</td>
<td>N</td>
<td>Y/N</td>
<td></td>
</tr>
<tr>
<td>using categories</td>
<td>N</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>N</td>
</tr>
<tr>
<td>using term absence</td>
<td>N</td>
<td>Y</td>
<td>N</td>
<td>N</td>
<td>N</td>
</tr>
<tr>
<td>performance in kNN/LLSF</td>
<td>excellent</td>
<td>excellent</td>
<td>excellent</td>
<td>poor</td>
<td>ok</td>
</tr>
</tbody>
</table>