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

Height	Eyes	Class
short	brown	hacker
tall	blue	theoretician
tall	brown	hacker
short	blue	theoretician

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

Data

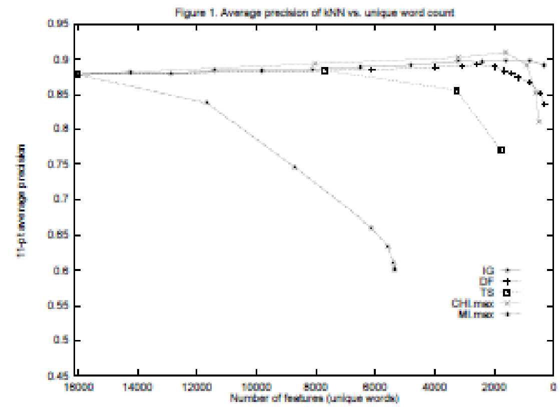
- OHSUMED
  - Subset of MEDLINE
    - 14,321 categories
    - 1990 abstracts: training
      - 72,076 terms
    - 1991 abstracts: testing
    - Average of 12 categories per document
- Evaluation
  - Recall
  - Precision
  - 11 point average precision

Term weighting

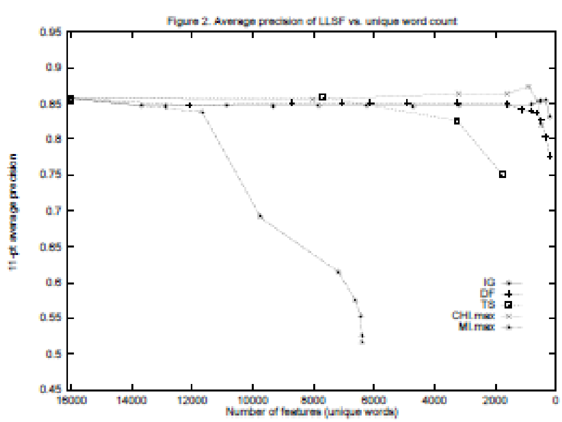
Term Frequency		Inverse Document Frequency		Normalization	
First Letter	$f(tf)$	Second Letter	$f(\frac{1}{df})$	Third Letter	$f(length)$
n (natural)	tf	n (no)	1	n (no)	1
l (logarithmic)	1+log(tf)	t (full)	$log(\frac{N}{df})$	c (cosine)	$\sqrt{w_1{}^2 + w_2{}^2 + \ldots + w_n{}^2}$
a (augmented)	$0.5 + 0.5 \times \frac{tf}{max\ tf}$				

Table 1: Term Weights in the Smart System

Performance curve: k-NN



Performance curve: LLSF



Qualitative comparison

Table 1. Criteria and performance of feature selection methods in kNN & LLSF

Method	DF	IG	CHI	MI	TS
favoring common terms	Y	Y	Y	N	Y/N
using categories	N	Y	Y	Y	N
using term absence	N	Y	Y	N	N
performance in kNN/LLSF	excellent	excellent	excellent	poor	ok