CS630 Representing and Accessing Digital Information

Text Classification: KNN and Rocchio

Thorsten Joachims
Cornell University

Test Collections

- **Reuters-21578**
  - Reuters newswire articles classified by topic
  - 90 categories (multi-label)
  - 9603 training documents / 3299 test documents (ModApte)
  - ~27,000 features
- **WebKB Collection**
  - WWW pages classified by function (e.g. personal HP, project HP)
  - 4 categories (multi-class)
  - 4183 training documents / 226 test documents
  - ~38,000 features
- **Ohsumed MeSH**
  - Medical abstracts classified by subject heading
  - 20 categories from "disease" subtree (multi-label)
  - 10,000 training documents/ 10,000 test documents
  - ~38,000 features

Example: Reuters Article (Multi-Label)

Categories: COFFEE, CRUDE

KENYAN ECONOMY FACES PROBLEMS, PRESIDENT SAYS

The Kenyan economy is heading for difficult times after a boom last year, and the country must tighten its belt to prevent the balance of payments swinging too far into deficit, President Daniel Arap Moi said.

In a speech at the state opening of parliament, Moi said high coffee prices and cheap oil in 1986 led to economic growth of five pct, compared with 4.1 pct in 1985. The same factors produced a two billion shilling balance of payments surplus and inflation fell to 5.6 pct from 10.7 pct in 1985, he added.

"But both these factors are no longer in our favour ... As a result, we cannot expect an increase in foreign exchange reserves during the year," he said.

…

Example: Ohsumed Abstract

Categories: Animal, Blood_Proteins/Metabolism, DNA/Drug_Effects, Mycotoxins/Toxicity, …

How aspartame prevents the toxicity of ochratoxin A.

Creppy EE, Baudrimont I, Anne-Marie

Toxicology Department, University of Bordeaux, France

The ubiquitous mycotoxin ochratoxin A (OTA) is found as a frequent contaminant of a large variety of food and feed and beverage such as beer, coffee and win. It is produced as a secondary metabolite of moulds from Aspergillus and Penicillium genera. Ochratoxin A has been shown experimentally to inhibit protein synthesis by competition with phenylalanine its structural analogue and also to enhance oxygen reactive radicals production. The combination of these basic mechanisms with the unusual long plasma half-life time (35 days in non-human primates and in humans), the metabolisation of OTA into still active derivatives and glutathione conjugate both potentially reactive with cellular macromolecules including DNA could explain the multiple toxic effects, cytotoxicity, teratogenicity, genotoxicity, mutagenicity and carcinogenicity. A relation was first recognised between exposure to OTA in the Balkan geographical

Multi-Class / Multi-Label

- **Cannot learn multi-label rules directly**
  - Most classifiers assume that each document is in exactly one class
  - Many classifiers can only learn binary classification rules
- **Most common solution: Multi-Label**
  - Learn one binary classifier for each label
  - Attach all labels, for which some classifier says positive
- **Most common solution: Multi-Class**
  - Learn one binary classifier for each label
  - Put example into the class with the highest probability (or some approximation thereof)

Performance Measures

- **Precision/Recall Break-Even Point**
  - Intersection of PR-curve with the identity line
- **Macro-averaging**
  - First compute the measure, then compute average
  - Results in average over tasks
- **Micro-averaging**
  - First average the elements of the contingency table, then compute the measure
  - Results in average over each individual classification decision

Experimental Results

Reuters Newswir e
- 90 cate gories
- 9603 training doc.
- 3299 test doc.
- ~27000 features

WebKB Collection
- 4 cate gories
- 4183 training doc.
- 226 test doc.
- ~38000 features

Ohsumed MeSH
- 20 cate gories
- 10000 training doc.
- 10000 test doc.
- ~38000 features

microaveraged precision/recall breakeven-point [0..100]	**Reuters**	**WebKB**	**Ohsumed**
Naive Bayes	72.3	82.0	62.4
Rocchio Algorithm	79.9	74.1	61.5
C4.5 Decision Tree	79.4	79.1	56.7
k-Nearest Neighbors	82.6	80.5	63.4
SVM	**87.5**	**90.3**	**71.6**

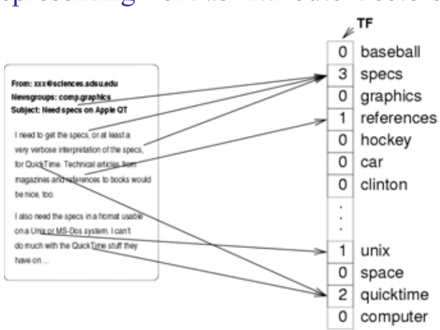
Rocchio Algorithm (Learning)

- **Given:** $(\vec{x}_1, y_1), \ldots, (\vec{x}_n, y_n) \sim P(X, Y)$
- **Preprocessing:**
  - Bring into vector space model representation (e.g. TFIDF)
  - Vectors normalized to Euclidian length 1
  - Split into set of positive / negative examples (ie. D_+ / D_-)
- **Training:**
  - Build prototype vector for each class
  - Compute weight vector as weighted difference between prototypes

$$\vec{w} = \frac{1}{|D_+|} \sum_{\vec{x}_i \in D_+} \vec{x}_i - \beta \frac{1}{|D_-|} \sum_{\vec{x}_i \in D_-} \vec{x}_i$$

  - Often: set negative elements of w vector to zero

Representing Text as Attribute Vectors



=> **Ignore ordering of words**

Rocchio Algorithm (Prediction)

- **Compute cosine between weight vector w and new example x'**
- **Prediction rule**

$$h(\vec{x}') = \begin{cases} 1 & \text{if } cos(\vec{w}, \vec{x}') > \theta \\ -1 & \text{else} \end{cases}$$

- **Threshold is a parameter, or often the cosine is just used to get a ranking**

K-Nearest Neighbor

- **Given:** $(\vec{x}_1, y_1), \ldots, (\vec{x}_n, y_n) \sim P(X, Y)$
- **Preprocessing:**
  - Bring into vector space model representation (e.g. TFIDF)
- **Learning:**
  - None
- **Prediction rule**

$$h(\vec{x}') = sign\left(\sum_{i \in knn(\vec{x}')} y_i cos(\vec{x}_i, \vec{x}') \right)$$

Experimental Results

Reuters Newswir e
- 90 cate gories
- 9603 training doc.
- 3299 test doc.
- ~27000 features

WebKB Collection
- 4 cate gories
- 4183 training doc.
- 226 test doc.
- ~38000 features

Ohsumed MeSH
- 20 cate gories
- 10000 training doc.
- 10000 test doc.
- ~38000 features

microaveraged precision/recall breakeven-point [0..100]	**Reuters**	**WebKB**	**Ohsumed**
Naive Bayes	72.3	82.0	62.4
Rocchio Algorithm	79.9	74.1	61.5
C4.5 Decision Tree	79.4	79.1	56.7
k-Nearest Neighbors	82.6	80.5	63.4
SVM	**87.5**	**90.3**	**71.6**

Feature (Subset) Selection

- **Some classifiers perform worse when using all features**
  - E.g. K-NN, Rocchio, C4.5, sometimes Naïve Bayes
- **Some classifiers are too inefficient to use all features**
  - E.g. C4.5
- **Methods**
  - Document Frequency Thresholding: Use only those words that occur at least *m* times in the training documents
  - Empirical Mutual Information: Pick words *w* with largest

$$I(w,y) = \sum_{y \in \{+1,-1\}} \sum_{w \in \{0,1\}} P(y,w) \log\left(\frac{P(y,w)}{P(y)P(w)}\right)$$

  - Also odds-ratio, chi-square score, stopword-removal, stemming

Comparison of Methods

	Naïve Bayes	Rocchio	C4.5	K-NN	SVM
Simplicity (conceptual)	+	++	-	++	-
Efficiency at training	+	+	--	++	-
Efficiency at prediction	++	++	+	--	++
Handling many classes	+	+	--	++	-
Theoretical understanding	o	--	-	o	+
Prediction accuracy	-	o	-	+	++
Stability and robustness	-	-	--	+	++