Inductive Learning Algorithms and Representations for Text Categorization

Susan Dumais John Platt David Heckerman

Mehran Sahami

Presenter: Haoran Hou

Text Categorization

real-time sorting emails/files

topic identification

structured search and/or browsing

finding documents that match long-term standing interests

Old School

Dewey Decimal

MeSH(Medical Subject Headings)

Yahoo!’s topic hierarchy

CyberPatrol

Inductive Learning Methods

Evaluation

Results & Others

Data:

a collection of hand-tagged financial newswire stories from Reuters.

http://www.research.att.com/~lewis/reuters21578.html (no longer available)

Inductive Learning Methods

Classifiers

Inductive Learning of Classifiers

Classifiers

$\rightarrow\top x = (x1,x2,x3\ldots xn)$

$f(\rightarrow\top x) = \textit{confidence(class)}$

eg. class- *interest*

if (interest AND rate) OR (quarterly), then confidence(cat interest) = 0.9

confidence(interest cat) = 0.3*interest + 0.4*rate + 0.7*quarterly

Inductive Learning Methods

Inductive Learning of Classifiers

Find Similar (a variant of Rocchio's method for relevance feedback)

Decision Tree

Naive Bayes

Naive Nets

SVM

*All methods require only on a small amount of labeled training data

The effectiveness of the model is tested on previously unseen instances.

Inductive Learning Methods

Inductive Learning of Classifiers

Find Similar (a variant of Rocchio's method for relevance feedback)

-tf*idf

-all features used

$$x_j = \alpha \cdot x_{q,j} + \beta \cdot \frac{\sum_{i \in rel} x_{i,j}}{n_r} + \gamma \cdot \frac{\sum_{i \in non-rel} x_{i,j}}{N - n_r}$$

$$= \cdot\beta \cdot \frac{\sum_{i \in rel} x_{i,j}}{n_r}$$

*no error minimization is applied

Inductive Learning Methods

Inductive Learning of Classifiers

Feature selection

$$MI(x_i, c) = \sum_{x_i \in \{0,1\}} \sum_{c \in \{0,1\}} P(x_i, c) \log \frac{P(x_i, c)}{P(x_i) P(c)}$$

SVM: K = 300

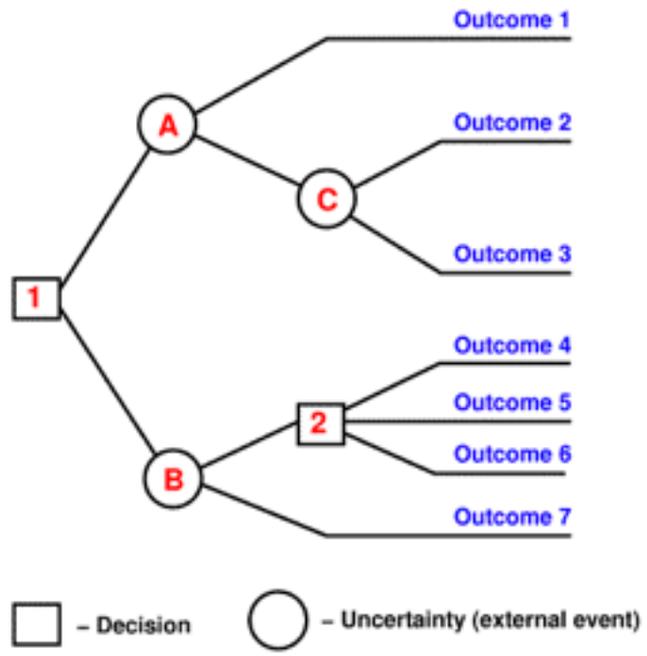
The remaining: K = 50

Only binary feature values are used

Inductive Learning Methods

Inductive Learning of Classifiers

Decision Tree



Recursive greedy splitting

Bayesian posterior probability

Node → class probability

Inductive Learning Methods

Inductive Learning of Classifiers

Naive Bayes

$$P(C = c_k \mid \vec{x}) = \frac{P(\vec{x} \mid C = c_k)P(C = c_k)}{P(\vec{x})}$$

Assume the features X1,….Xn are conditionally independent

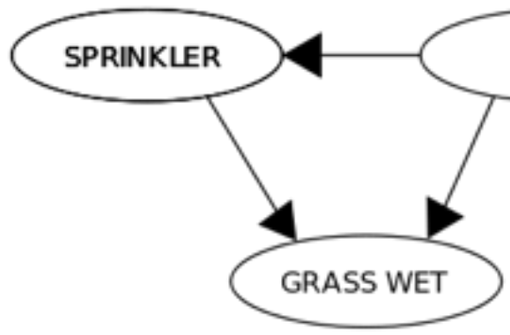
$$P(\vec{x} \mid C = c_k) = \prod_i P(x_i \mid C = c_k)$$

Inductive Learning Methods

Inductive Learning of Classifiers

Bayes Nets

RAIN	SPRINKLER T	SPRINKLER F
F	0.4	0.6
T	0.01	0.99



RAIN T	RAIN F
0.2	0.8

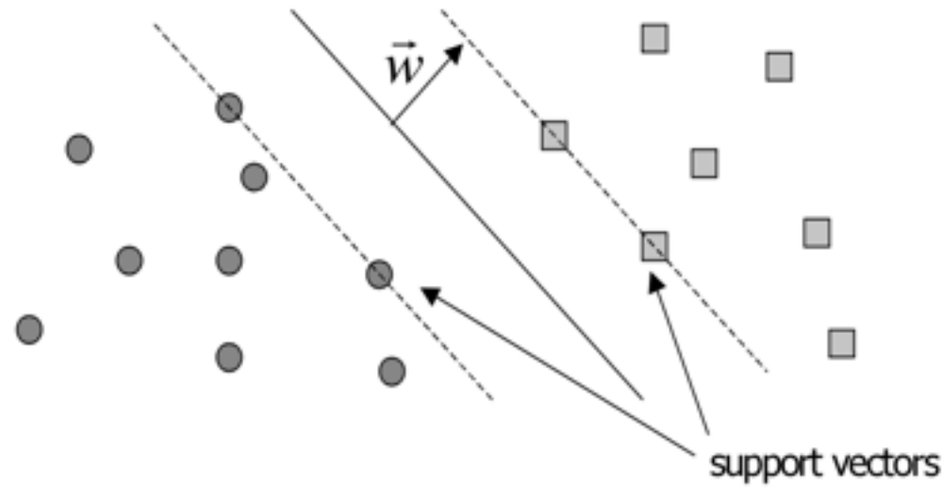
SPRINKLER	RAIN	GRASS WET T	GRASS WET F
F	F	0.0	1.0
F	T	0.8	0.2
T	F	0.9	0.1
T	T	0.99	0.01

2-dependence Bayesian classifier

Inductive Learning Methods

Inductive Learning of Classifiers

SVM



Simplest linear version

$$y_i(\vec{w} \cdot \vec{x}_i - b) \geq 1$$

Inductive something something ⟶ Evaluation

Evaluation

Reuters-21578

Summary of Inductive Learning Process

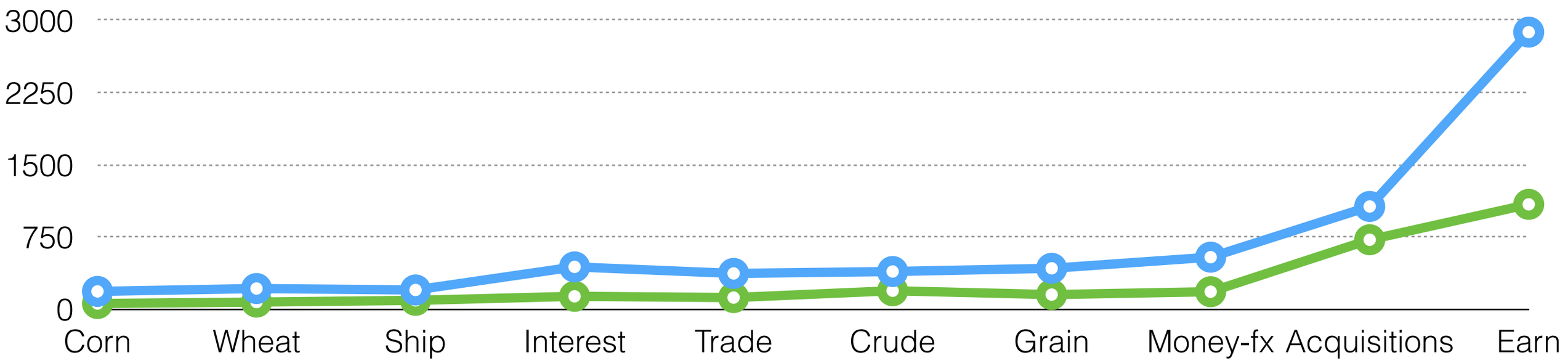
Inductive something something → Evaluation

Reuters-21578

21578 collection, 200 words in length

118 categories

75% train, 25% test

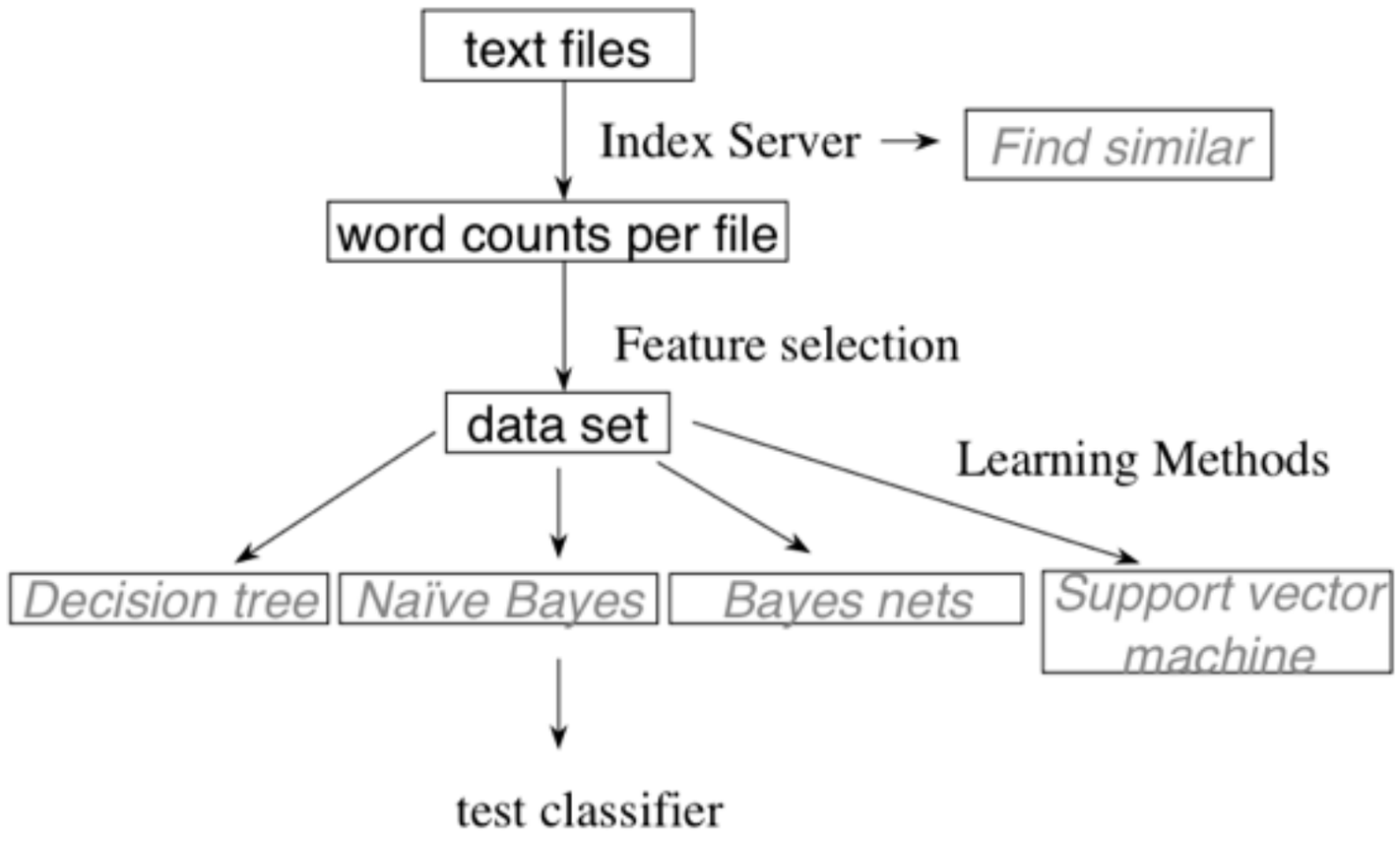


Inductive something something ⟶ Evaluation

Category Name	Num Train	Num Test
Earn	2877	1087
Acquisitions	1650	719
Money-fx	538	179
Grain	433	149
Crude	389	189
Trade	369	118
Interest	347	131
Ship	197	89
Wheat	212	71
Corn	182	56

Inductive something something ⟶ Evaluation

Summary of Inductive Learning Process



Average of precision and recall(F measure?)

Train/test dataset not optimized

Something something something → Evaluation → Results

Results & Others

Training Time

Classification Speed for New Instances

Classification Accuracy

Other Experiments

Inductive something something → Evaluation → Results

Training Time

266 MHz Pentium II running Windows NT.

Fastest: Find Similar (<1 CUP sec/cat)

SVM (<2 CUP sec/cat)

Naive Bayes(8 CPU sec/cat)

Decision Trees (~70 CUP sec/cat)

Slowest: Bayes Nets(~145 CUP sec/cat)

Inductive something something → Evaluation → Results

	Findsim	NBayes	BayesNets	Trees	LinearSVM
earn	92.9%	95.9%	95.8%	97.8%	98.0%
acq	64.7%	87.8%	88.3%	89.7%	93.6%
money-fx	46.7%	56.6%	58.8%	66.2%	74.5%
grain	67.5%	78.8%	81.4%	85.0%	94.6%
crude	70.1%	79.5%	79.6%	85.0%	88.9%
trade	65.1%	63.9%	69.0%	72.5%	75.9%
interest	63.4%	64.9%	71.3%	67.1%	77.7%
ship	49.2%	85.4%	84.4%	74.2%	85.6%
wheat	68.9%	69.7%	82.7%	92.5%	91.8%
corn	48.2%	65.3%	76.4%	91.8%	90.3%
Avg Top 10	64.6%	81.5%	85.0%	88.4%	92.0%
Avg All Cat	61.7%	75.2%	80.0%	N/A	87.0%

Table 2 – Breakeven Performance for 10 Largest Categories, and over all 118 Categories.

Inductive something something → Evaluation → Results

New Instances?

All less than 2 sec

Inductive something something → Evaluation → Results

Accuracy

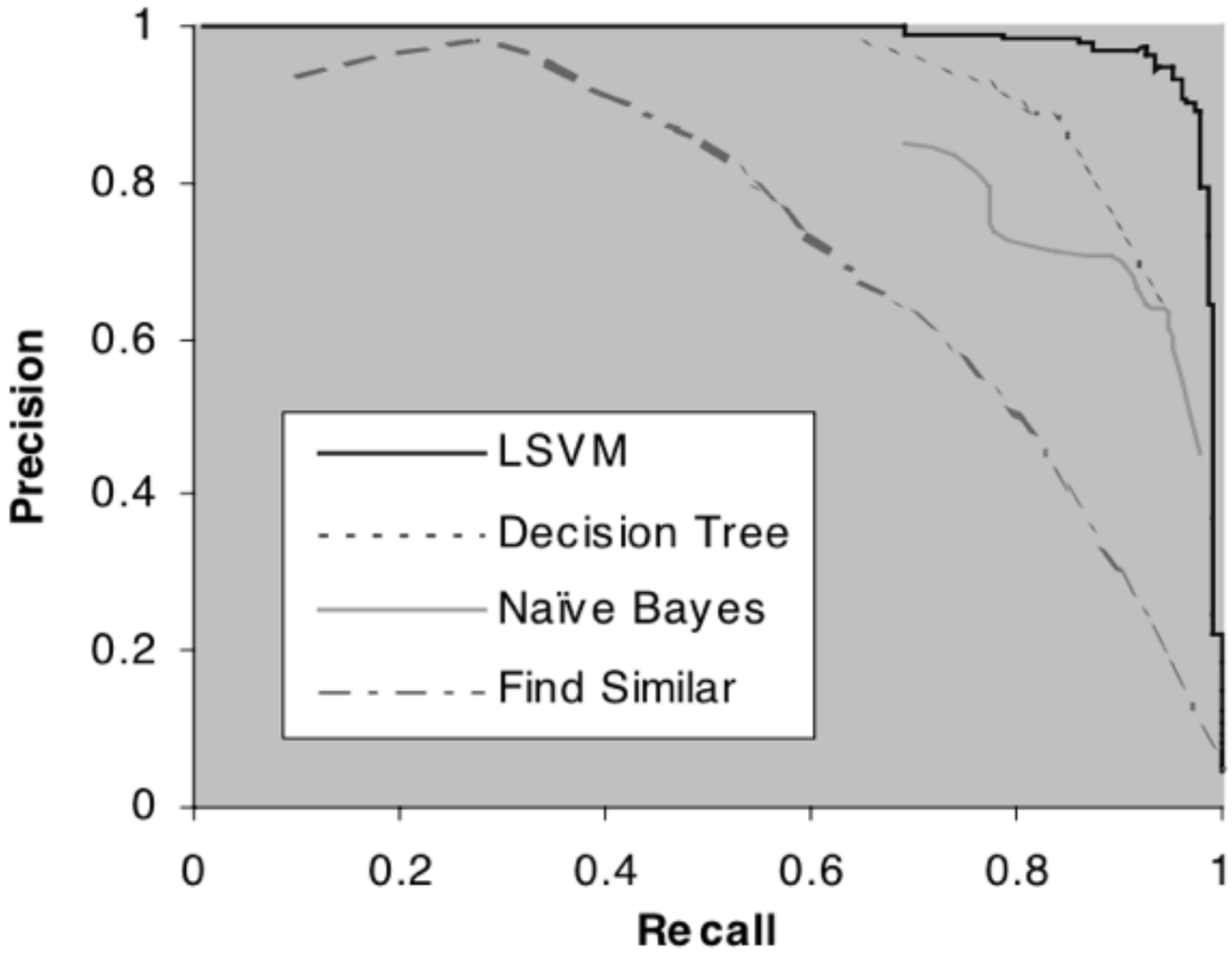


Figure 3 – Precision-Recall Curve for Category "grain"

Inductive something something → Evaluation → Results

Others?

Sample Size

N-gram

Binary vs. 0/1/2 features

Questions?