A Continuum Between Discriminant and Model-Based Classifiers

Jason Rennie
jrennie@ai.mit.edu

Joint work with Kai Shih and Yu-han Chang.

Text Classification

- Assign text document a label based on content.
- Important aspect of information management.
- Examples:
  - E-mail filtering
  - E-commerce

Example: E-mail Classification

- Filter e-mail into folders set up by user.
- Aids searching for old e-mails
- Can be used to prioritize incoming e-mails
  - High priority to e-mails concerning your Ph.D. thesis
  - Low priority to “FREE Pre-Built Home Business”

Example: E-Commerce

- Users locate products in two basic ways: search and browsing.
- Browsing is best when user doesn't know exactly what he/she wants.
- Text classification can be used to organize products into a hierarchy according to description.
- EBay: Classification can be used to ensure that product fits category given by user.

Representation

```
From: dyer@spdcc.com (Steve Dyer)
Subject: Re: food-related seizures?

My comments about the Feingold Diet have no relevance to your
daughter's purported FrostedFlakes-related seizures.  I can't imagine
why you included it.
```

$\downarrow$

food	1
seizures	2
diet	1
catering	0
religion	0
$\vdots$	$\vdots$

Speed is Important!

- Training examples in 1000's
- High dimensionality (vocabularies of 10,000+)
- Slow training ($O(n^2)$, $O(n^3)$) is impractical
- Real applications are especially time critical (imagine waiting 5 minutes while the system corrects for a misclassified e-mail!)

Can We Trade Speed for Accuracy?

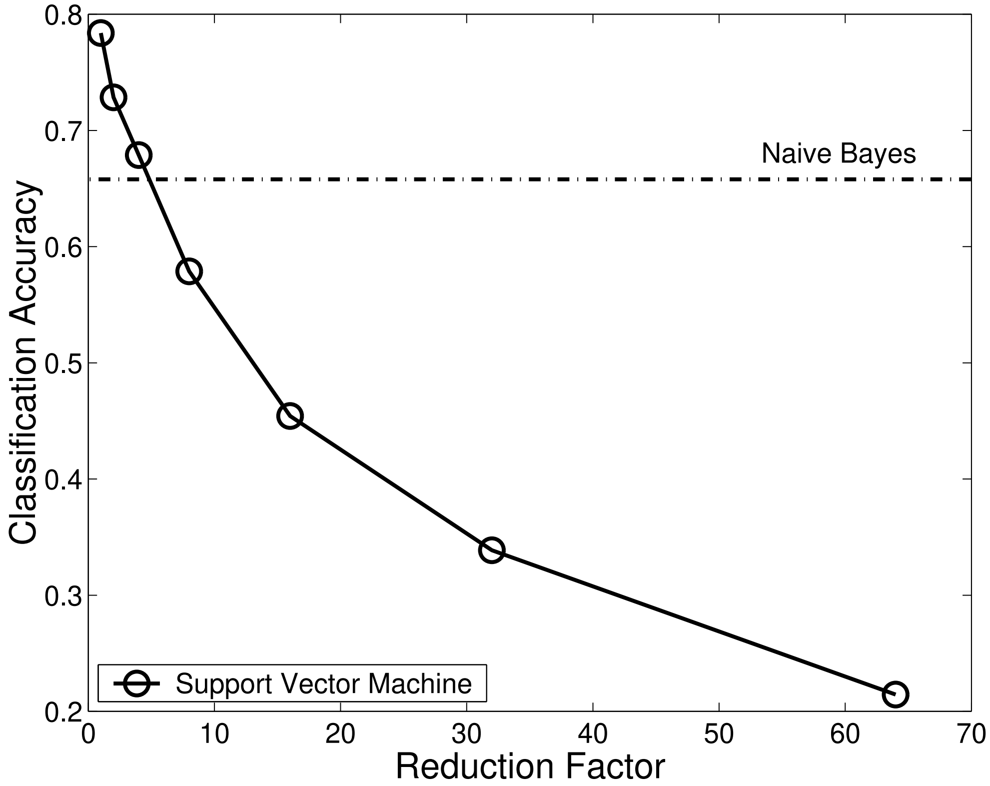
- Not easily
- Most algorithms are either "fast" or "slow"

Fast (Linear)	**Slow** (Low order polynomial)
Naive Bayes	Support Vector Machine
Rocchio	k-Nearest Neighbor
	Neural Network
	Least-squares fit

- Currently no bridge exists

Subsampling

- Subsampling eliminates some training documents to make training faster.
- But, accuracy suffers...

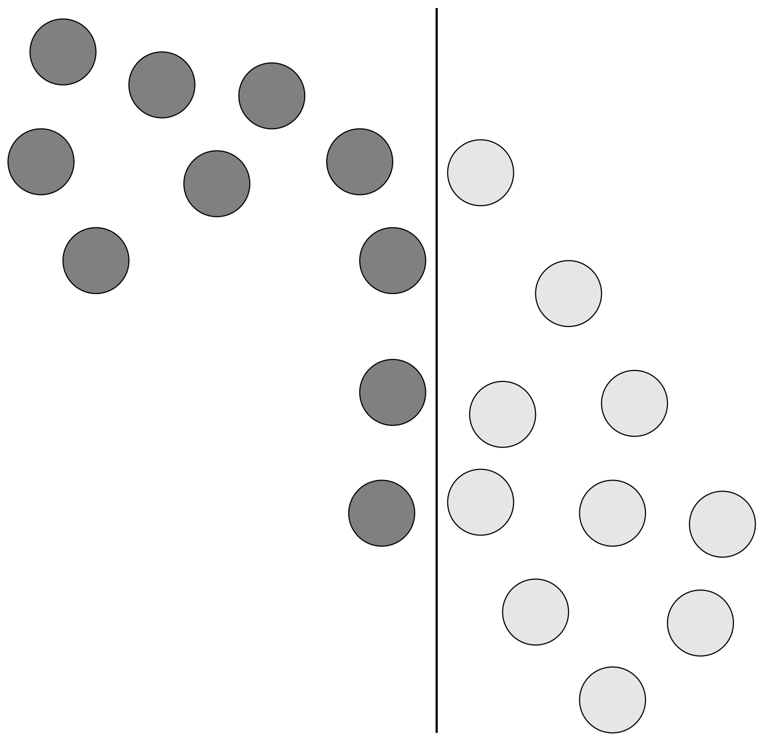


Bundled-SVM: Preserve Sufficient Statistics

- Fast classification algorithms (e.g. Naive Bayes, Rocchio) are based on mean
- Idea: reduce training data but preserve sufficient statistics
- Sufficient statistics are enough for training of simple algorithms
- Worst we should do is as well as simple algorithms

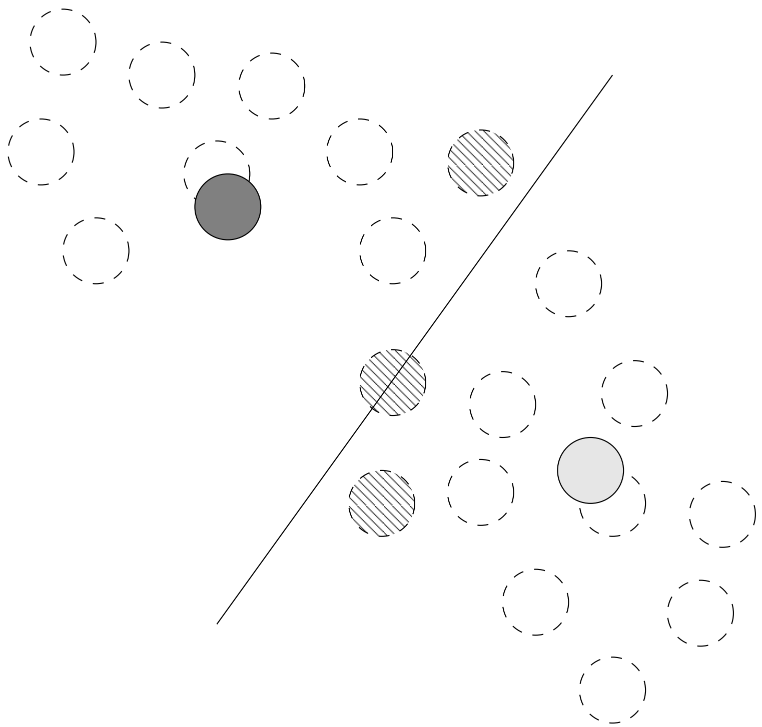
Accurate Endpoint: SVM

- Support Vector Machine uses points on boundary of two classes to determine decision plane.
- Training time is approximately quadratic.



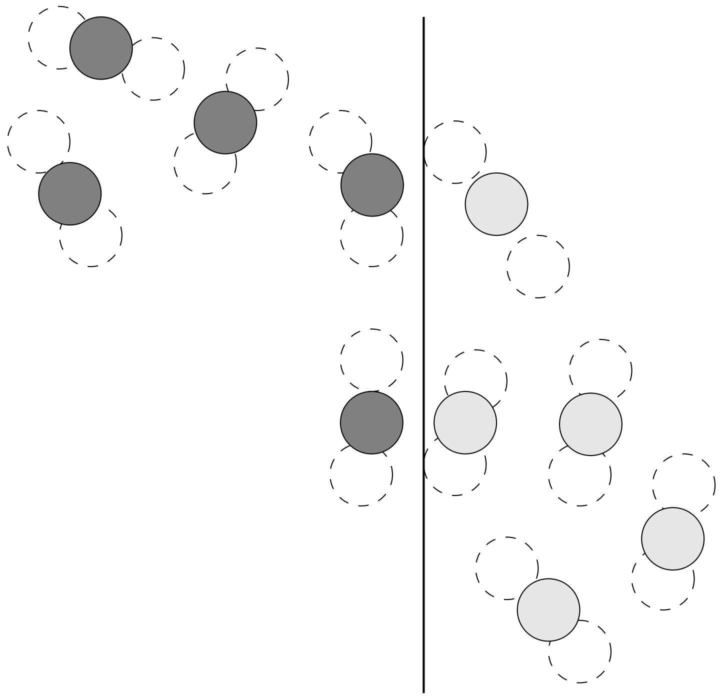
Fast Endpoint: Rocchio/Naive Bayes

- Rocchio and Naive Bayes choose a boundary based on the mean of the data.
- Training time is linear.

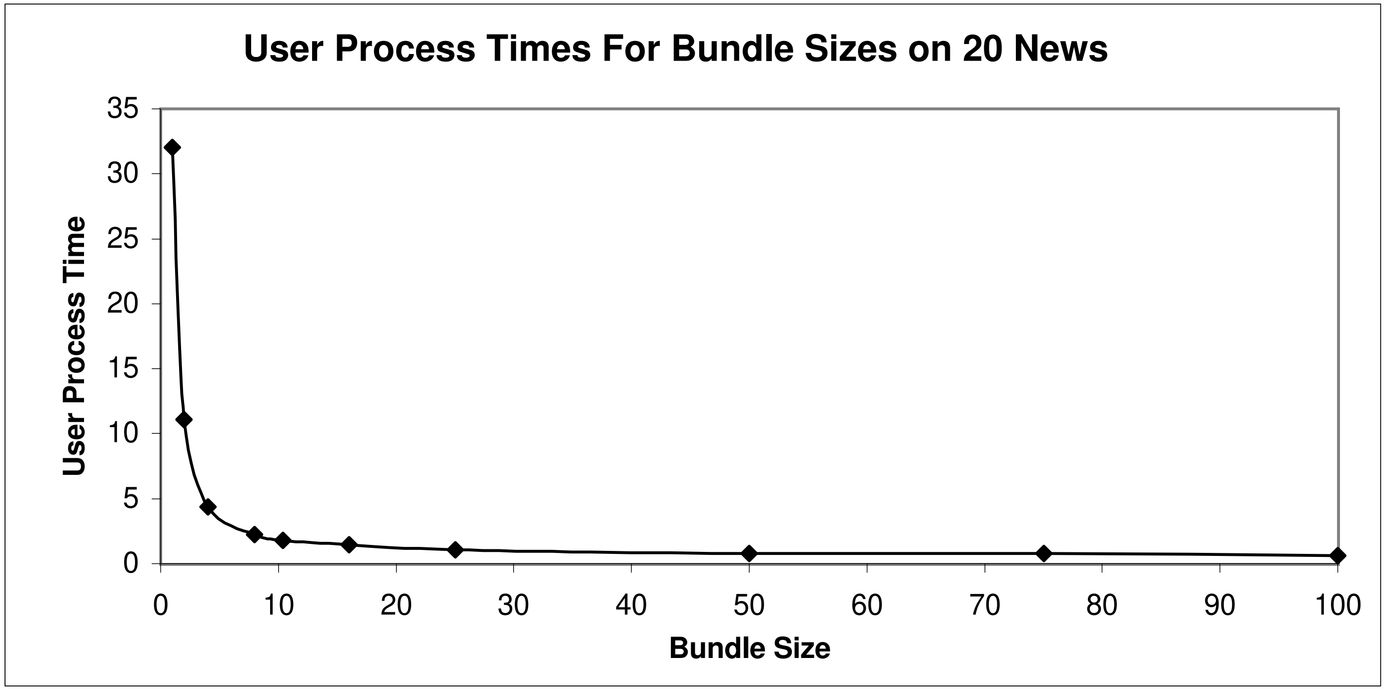


Bundled-SVM

- Feed bundled (averaged) data points to SVM.
- Time/accuracy trade-off based on amount of bundling.

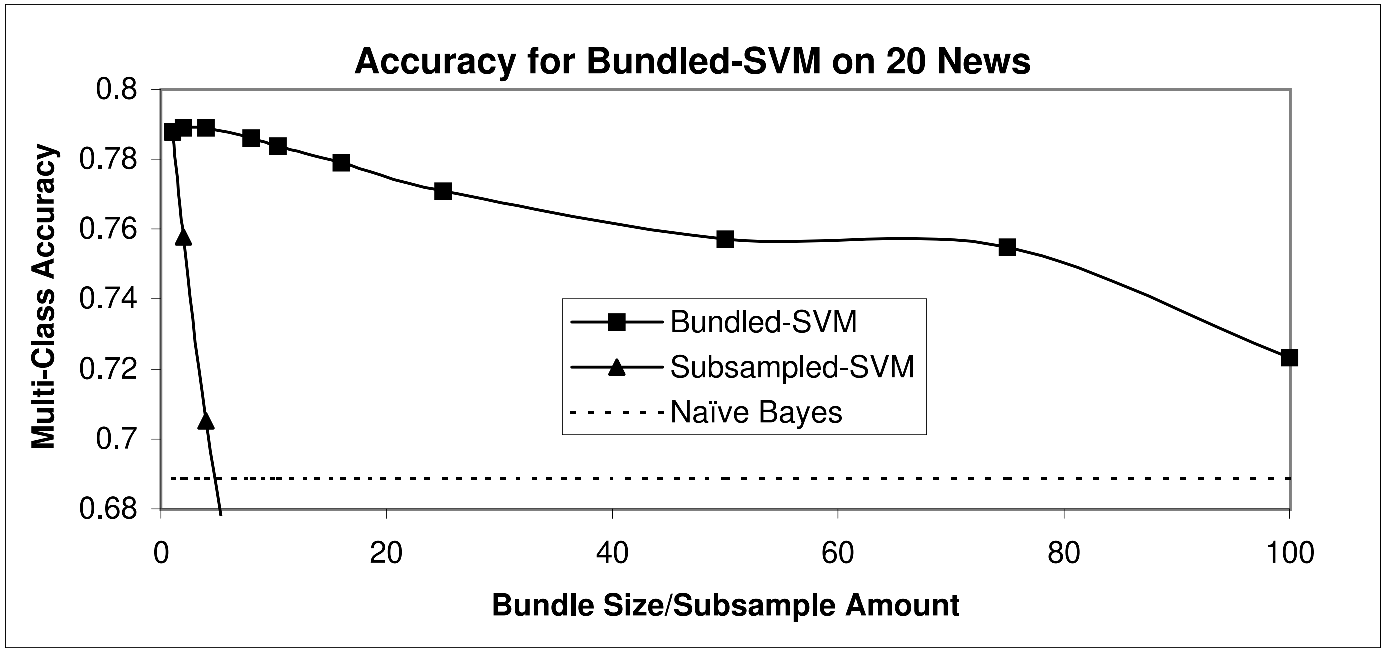


The Bundled-SVM is very fast



The Bundled-SVM is very accurate

- The Bundled-SVM creates a *continuum* of classifiers that trade off speed for accuracy.



Classification Results

- Accuracy never decreases below Naive Bayes baseline

Data Set	SVM	$t = 2$	$t = \sqrt{n}$	$t = \min$	NB
Reuters (micro)	0.857	0.884	0.857	0.708	0.736
Reuters (macro)	0.631	0.681	0.586	0.508	0.264
20 News	0.788	0.788	0.784	0.723	0.689
Industry Sector	0.928	0.909	0.878	0.901	0.566

Conclusions

- The Bundled-SVM is an effective way to trade off speed for accuracy
- It speeds up training by using a reduced data set
- Unlike subsampling, it preserves the important sufficient statistics of the data
- Bundling algorithm works for arbitrary mean statistics (e.g. Gaussian with full covariance matrix)