Combining Rankers Using Conditional Models on Permutations

Guy Lebanon & **John Lafferty**
School of Computer Science
Carnegie Mellon University

May 20, 2002

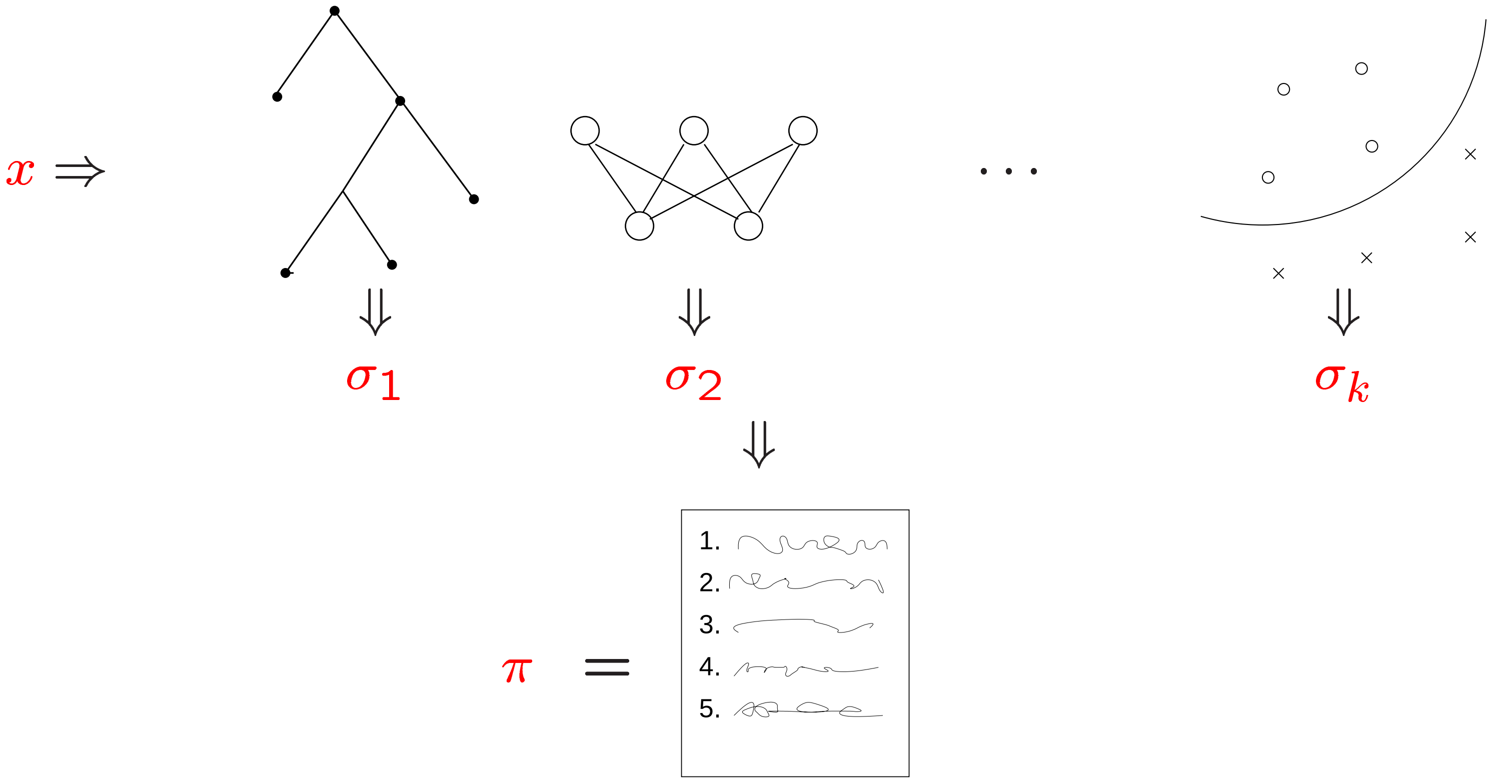
Motivation

Task: Calculate a (partial) ranking π of items $y_1, \ldots, y_n$ given several other rankings $\{\sigma_i\}$ of $y_1, \ldots, y_n$

Examples:

- In classification $y_1, \ldots, y_n$ are labels and $\{\sigma_i\}$ are features or sufficient statistics.
- In Meta search, $y_1, \ldots, y_n$ are web pages and $\{\sigma_i\}$ are web-page rankings given by other search engines.

Ensemble Classification as Ranking



Combined ranking

Meta-Search as Ranking

- Previous work on meta-search using machine learning:
  RankBoost (Freund et al. 1998)
  Bayes Fuse (Aslam & Montague, 2001)

- Modeling relevance scores (Manmatha et al. 2001)

- Most prior work in information retrieval uses linear combination of scores.

- Our approach is to estimate models on *rankings*

We do not assume that we have *score* information thereby by-passing

- The problem of non-calibrated features
  e.g. kNN, SVM, neural net

- The need for scores
  e.g. commercial search engines

Outline

- Algebraic Concepts
- Mallows Model
- Conditional Model for Combining Rankers
- Extensions
- Parameter Estimation and Inference
- Experiments

Algebraic Concepts - Permutations

- A permutation π is a bijection from $\{1, \ldots, n\}$ to $\{1, \ldots, n\}$
  $\pi(1) = 2 \quad \pi(2) = 1 \quad \pi(3) = 3$
  $e(1) = 1 \quad e(2) = 2 \quad e(3) = 3$

- The set of all permutations together with function composition operation forms the symmetric group $\mathfrak{S}_n$.
  $\pi^{-1}(1) = 2 \quad \pi^{-1}(2) = 1 \quad \pi^{-1}(3) = 3$
  $\pi\pi^{-1} = \pi^{-1}\pi = e$

- A permutation corresponds to a full ranking over the items $y_1, \ldots, y_n$.

Algebraic Concepts - Cosets I

- Let $\mathfrak{S}_{n-k}$ be the set of permutations in $\mathfrak{S}_n$ that fix the first k elements

$$\mathfrak{S}_{n-k} = \{\sigma \in \mathfrak{S}_n : \sigma(1) = 1, \ldots, \sigma(k) = k\}$$

- A coset $\mathfrak{S}_{n-k}\pi$ is the set of permutations $\{\sigma\pi : \sigma \in \mathfrak{S}_{n-k}\}$.

- The set of all such cosets forms the quotient group $\mathfrak{S}_n/\mathfrak{S}_{n-k}$.

- Such a coset corresponds to a ranking of the top k items

$$\pi(2) = 1, \pi(4) = 2, \quad \mathfrak{S}_{5-2}\pi : \langle 2, 4, *, *, * \rangle$$

Algebraic Concepts - Cosets II

- Let $\mathfrak{S}_\lambda \cong \mathfrak{S}_{n-n_1} \times \cdots \times \mathfrak{S}_{n-n_k}$ be the set of permutations in $\mathfrak{S}_n$ that permute only within the $\{n_i\}$ sets
  E.g. $\mathfrak{S}_{5-3} \times \mathfrak{S}_{5-2}$ permutes the top 3 elements and the bottom 2 elements

$$\begin{aligned} & \sigma \in \mathfrak{S}_{5-3} \times \mathfrak{S}_{5-2} \\ \Rightarrow \quad & \sigma(\{1,2,3\}) = \{1,2,3\}, \quad \sigma(\{4,5\}) = \{4,5\} \end{aligned}$$

- Such cosets correspond to partial rankings

$$\mathfrak{S}_{5-3} \times \mathfrak{S}_{5-2}\ \pi\ =\ \{1,3,5\} \big| \{2,4\}$$

Distances on Permutations

A distance on permutations is a function $d : \mathfrak{S}_n \times \mathfrak{S}_n \to \mathbb{R}$ that satisfies

$$\begin{aligned}
d(\pi, \pi) &= 0, & \forall \pi \in \mathfrak{S}_n \\
d(\pi, \sigma) &> 0, & \forall \pi, \sigma \in \mathfrak{S}_n, \quad \pi \neq \sigma \\
d(\pi, \sigma) &= d(\sigma, \pi), & \forall \pi, \sigma \in \mathfrak{S}_n \\
d(\pi, \sigma) &\leq d(\pi, \tau) + d(\tau, \sigma), & \forall \pi, \sigma, \tau \in \mathfrak{S}_n
\end{aligned}$$

It also makes sense to require right-invariance to arbitrary relabeling of $y_1, \ldots, y_n$

$$d(\pi, \sigma) = d(\pi\tau, \sigma\tau) \quad \forall \pi, \sigma, \tau \in \mathfrak{S}_n.$$

Some Distances on Permutations

The following (right-invariant) distances are Spearman's rank correlation R and footrule F, and Kendall's τ

$$
\begin{aligned}
R^2(\pi,\sigma) &= \sum_{i=1}^{n} (\pi(i)-\sigma(i))^2 \\
F(\pi,\sigma) &= \sum_{i=1}^{n} |\pi(i)-\sigma(i)| \\
T(\pi,\sigma) &= \text{minimum number of adjacent transpositions} \\
&\qquad \text{needed to bring } \pi^{-1} \text{ to } \sigma^{-1}
\end{aligned}
$$

E.g. if $\pi^{-1} = \langle 1,2,3\rangle, \sigma^{-1} = \langle 3,1,2\rangle$ then $T(\pi,\sigma) = 2$

$$\langle 1,2,3\rangle \to \langle 1,3,2\rangle \to \langle 3,1,2\rangle$$

Mallows Model

Mallows model $\mathcal{M}_d(\mu, \theta)$ is the analog of a normal distribution over $\mathfrak{S}_n$

$$p(\pi; \mu, \theta) \;=\; \frac{1}{Z(\mu, \theta)} e^{\theta\, d(\pi,\mu)}$$

- $\mu \in \mathfrak{S}_n$ location parameter, $\theta \in \mathbb{R}$ scale parameter
- When using a right-invariant metric $Z(\mu, \theta) = Z(\theta)$
- Multistage extensions by Fligner and Verducci (1986, 1988)

Conditional Model for Combining Rankers

We extend Mallows model to the following *conditional* model (see also Feigin 1993)

$$p(\pi \mid \boldsymbol{\sigma}; \boldsymbol{\theta}) \;=\; \frac{1}{Z(\boldsymbol{\theta}, \boldsymbol{\sigma})}\, e^{\sum_j \theta_j d(\pi, \sigma_j)}$$

$\boldsymbol{\theta}$ are estimated using *conditional* maximum likelihood over the data $D = \{(\pi^{(i)}, \boldsymbol{\sigma}^{(i)})\}_i$

$$\boldsymbol{\theta} = \arg\max_{\boldsymbol{\theta}} \; \log \prod_i p(\pi^{(i)} | \boldsymbol{\sigma}^{(i)}; \boldsymbol{\theta})$$

Extensions

- Non uniform carrier density

$$p(\pi \,|\, \boldsymbol{\sigma}; \boldsymbol{\theta}) \;=\; \frac{1}{Z}\, p(\pi)\, e^{\sum_j \theta_j d(\pi,\sigma_j)}$$

- Ranker Interaction Terms

$$p(\pi \,|\, \boldsymbol{\sigma}; \boldsymbol{\theta}) \;=\; \frac{1}{Z}\, e^{\sum_j \theta_j d(\pi,\sigma_j) + \sum_{ij} \theta_{ij} d(\pi,\sigma_i) d(\pi,\sigma_j)}$$

- Ranker specialization, $\nu_i : \mathcal{X} \to \{0, 1\}$

$$p(\pi \,|\, \boldsymbol{\sigma}, \nu; \boldsymbol{\theta}) \;=\; \frac{1}{Z}\, e^{\sum_{ij} \nu_i \theta_{ij} d(\pi,\sigma_j)}$$

Bayesian Interpretation

Following Fligner & Verducci (1990) we can give the following interpretation

- Suppose σ_j are independently sampled from Mallows models

$$\sigma_j \sim \mathcal{M}_d(\pi, \theta_j)$$

- Then under a prior $p(\pi)$ the posterior for π is

$$p(\pi|\boldsymbol{\sigma}; \boldsymbol{\theta}) = \frac{1}{Z} p(\pi)\, e^{\sum_j \theta_j d(\pi, \sigma_j)}$$

Parameter Estimation

In practical situation relevance is often given in terms of cosets (partial order) e.g. classification $y^{(i)} \equiv \mathfrak{S}_{n-1}\pi^{(i)}$

Maximize marginal conditional likelihood (by conjugate gradient)

$$\ell(\theta) = \log \prod_i \sum_{\pi \in \mathfrak{S}_{n-1}\pi^{(i)}} p(\pi \,|\, \boldsymbol{\sigma}^{(i)}; \boldsymbol{\theta}) = \sum_i \log \left(\frac{\sum_{\pi \in \mathfrak{S}_{n-1}\pi^{(i)}} e^{\sum_j \theta_j d(\pi, \sigma_j^{(i)})}}{\sum_{\pi \in \mathfrak{S}_n} e^{\sum_j \theta_j d(\pi, \sigma_j^{(i)})}} \right)$$

If n large estimate the sums by MCMC

Inference

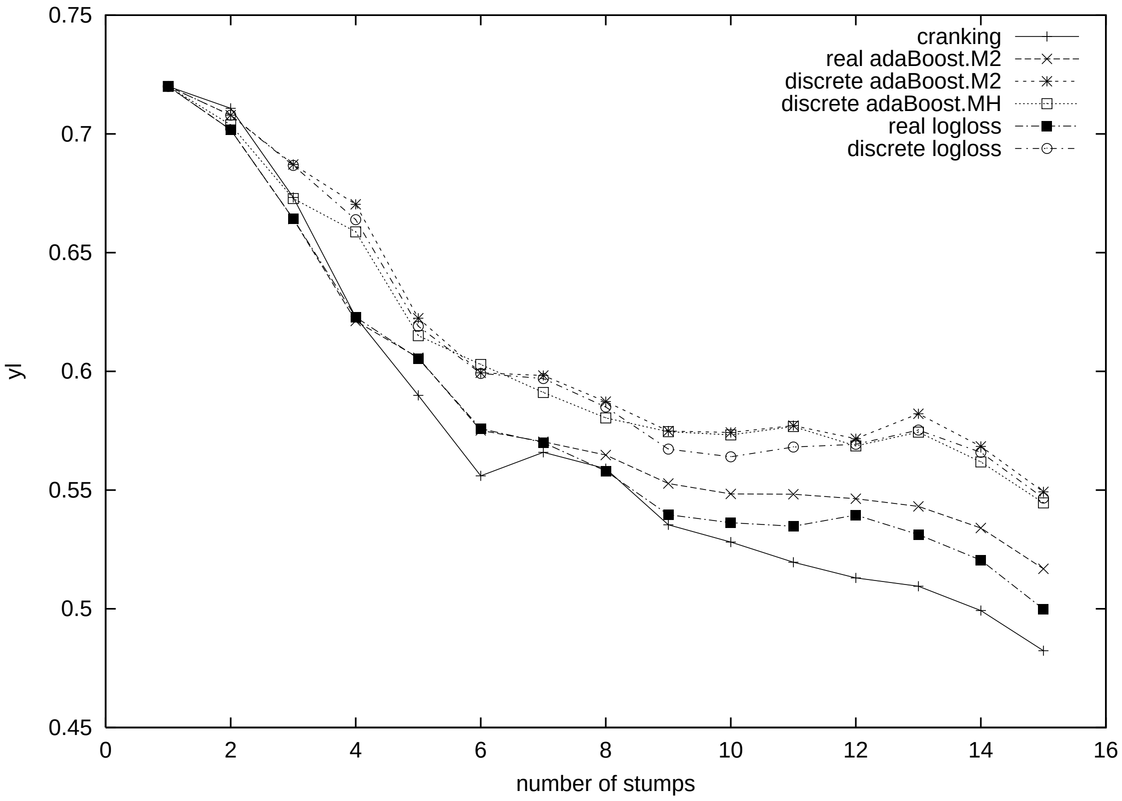
The model gives probabilities over full rankings of $y_1, \ldots, y_n$
Other options are:

- Rank $y_1, \ldots, y_n$ by the expected rank of each item

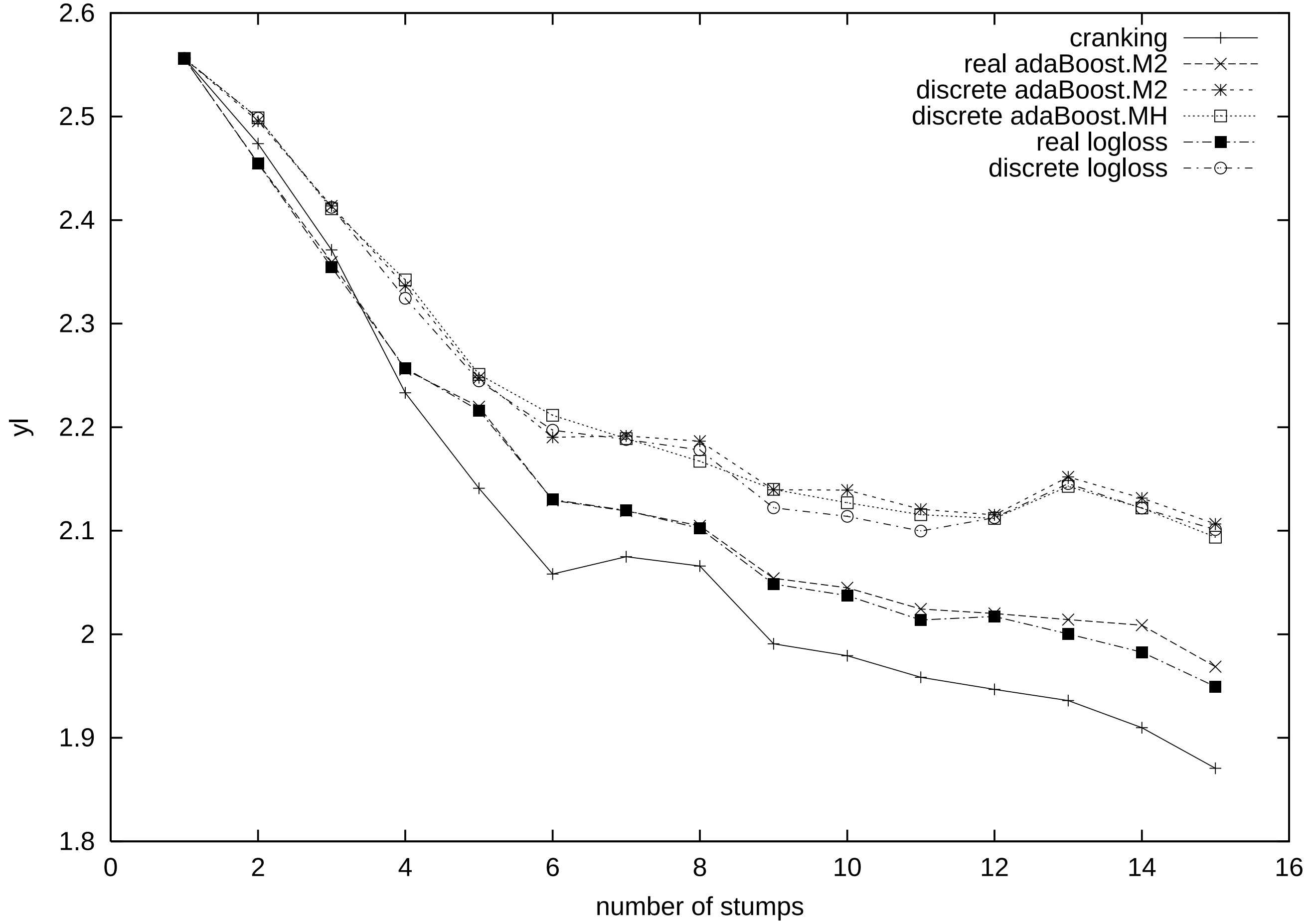
$$E\left[\pi(y^{(i)}) \,|\, \boldsymbol{\sigma}^{(i)}; \boldsymbol{\theta}\right] = \sum_{k=1}^{n} k\, p(\pi(y^{(i)})) = k \,|\, \boldsymbol{\sigma}^{(i)}; \boldsymbol{\theta})$$

- Sample from $p(\pi \,|\, \boldsymbol{\sigma}; \boldsymbol{\theta})$ and compute expectations of statistics.

Classification Experiment: Mixture of 5 Gaussians



Test error rate vs. number of (random) decision stumps



Test rank rate vs. number of (random) decision stumps

Metasearch Experiment

Data: Queries with names of ML researchers, expanded in different ways and used to retrieve Web pages (Cohen, Schapire & Singer, 1999)

- 14 input rankers
- Unique correct answer
- Conditional model is trained to maximize marginal likelihood, using a Metropolis-Hastings algorithm.

A Naïve Bayes Model

For comparison, we use the “Bayes Fuse” model (Aslam & Montague, 2001)

Rank documents according to posterior log-odds

$$\log\left(\frac{p(\mathrm{rel}\,|\,r_1,\ldots,r_k)}{p(\overline{\mathrm{rel}}\,|\,r_1,\ldots,r_k)}\right) \;=\; \log\left(\frac{p(\mathrm{rel})\prod_i p(r_i\,|\,\mathrm{rel})}{p(\overline{\mathrm{rel}})\prod_i p(r_i\,|\,\overline{\mathrm{rel}})}\right)$$

$p(r_i\,|\,\mathrm{rel})$ estimated using Dirichlet prior and kernel density estimate.

Coset size	*Bayes Fuse* (Laplace)	*Bayes Fuse* (Kernel)	*Cranking*	*Best ranker*
3	1.5	1.7	2.5	3.3
7	3.6	3.5	3.0	4.8
10	3.6	3.1	2.7	6.0

Permutation model gives better results when ranking a larger number of pages. More computationally expensive.

Summary

- Ranking models appropriate for many applications
- Bypass feature calibration and score unavailability problems
- Competitive with state-of-the-art classifiers and meta-search algorithms
- More computationally expensive
- Natural for multi-label classification
- Simple conditional model; many extensions possible

Acknowledgments

Thanks to

Joseph Verducci	for helpful comments
William Cohen	for providing the meta search data