Cranking: Combining Rankings Using Conditional Probability Models on Permutations

Guy Lebanon LEBANON@CS.CMU.EDU
John Lafferty LAFFERTY@CS.CMU.EDU
School of Computer Science, Carnegie Mellon University, Pittsburgh, PA 15213 USA

Abstract

A new approach to ensemble learning is introduced that takes ranking rather than classification as fundamental, leading to models on the symmetric group and its cosets. The approach uses a generalization of the Mallows model on permutations to combine multiple input rankings. Applications include the task of combining the output of multiple search engines and multiclass or multilabel classification, where a set of input classifiers is viewed as generating a ranking of class labels. Experiments for both types of applications are presented.

1. Introduction

Many machine learning problems involve the analysis of ranked data. As an example, in the information retrieval fusion problem, one is presented with a ranked list of Web pages output by various search engines, and the task is to somehow combine them to obtain a more accurate "meta-search" engine. The problem is particularly challenging since typically only the rankings are available, and not scores on individual items.

A seemingly unrelated problem is to combine classifiers using what is commonly referred to as an ensemble method. In an ensemble approach to classification, an input x receives a score for each candidate label y, according to functions $f_i(x, y)$ that are thought of as votes or confidences for assigning label y to instance x using the i-th classifier. Ensemble methods such as AdaBoost (Freund & Schapire, 1996) combine the classifiers using linear combinations of these scores. However, often the input classifiers do not have scores associated with them, or their scores may not be comparable or well calibrated. An alternative approach is to view each input classifier in terms of the ranked list of labels that it assigns to x. Under this view it is natural to build probability distributions over rankings of the labels, leading to models on permutation groups. This is a largely unexplored approach in machine learning, and the one that is pursued in this paper.

While there has been little previous work on models for ranked data in the machine learning literature, there is a significant body of work on such models in statistics. Much of this has focused on simple generative models, estimating a parametric distribution $p(\pi \mid \theta)$ where π is a permutation or coset, corresponding to a partial ranking. Early work in this direction includes the Thurstone model (Thurstone, 1927) and the Babington Smith model (Smith, 1950). Mallows (1957) proposed a metric-based unimodal distribution that is a special case of the Babington Smith model. Fligner and Verducci's multistage models (Fligner & Verducci, 1986; Fligner & Verducci, 1988) are a generalization of the Mallows model for multistage rankings. The use of group representations as a tool for approaching such problems has been championed by Diaconis, with emphasis on analysis of variance methods (Diaconis, 1988; Diaconis, 1989).

This paper explores conditional models on permutations as a tool for solving problems involving the analysis of ranked data, such as fusion or multiclass classification. The models are conditional because they take as input a set of permutations. Our most basic model is an extension of the Mallows model to the conditional setting. While some attempts have been made in the statistical literature to add covariates, they rarely take the form of additional rankings; see (Fligner & Verducci, 1993) for a recent collection of relevant papers. An interesting feature of the model proposed in this paper, as explained in detail below, is that because of the invariance properties of the sufficient statistics, the model has a natural Bayesian interpretation with respect to an underlying generative model. Thus, in contrast to many ensemble methods that are purely discriminative, the approach introduced here has both discriminative and generative interpretations.

We view this work as forming a bridge between the statistical literature on models for ranked data and the machine learning perspective on algorithms and architectures. The following sections present the new approach, which we call *cranking*, together with the results of experiments that validate it. Section 2 reviews the basic concepts that are needed from the theory of permutation groups. Section 3 presents the model that the approach is based upon, and various interpretations of this model are described in Section 3.3. Learning and inference are described in Section 4, followed by extensions to the model in Section 5. Section 6 presents experiments that were carried out on synthetic data, multiclass problems from the UCI repository, and actual meta-search data. The results of the paper are summarized in Section 7.

2. Metrics, Permutations and Coset Spaces

This section reviews some basic concepts from permutation theory that will be of use in later sections, adopting the notation and metrics of Critchlow (1980).

Let $\mathcal{X} = y_1, \dots, y_n$ be a set of items to be ranked, identified with the numbers $1, \dots, n$. A permutation π is a bijection from $\{1, \dots, n\}$ to itself. We use $\pi(i)$ to denote the rank given to item i and $\pi^{-1}(i)$ to denote the item assigned to rank i. We will usually write π and π^{-1} as vectors whose i-th component is $\pi(i)$ and $\pi^{-1}(i)$, respectively. The collection of all permutations of n-items forms a non-abelian group under composition, called the *symmetric group of order* n, and denoted $\mathcal{S}_n$.

We are given a set of training instances that we wish to assign a ranking of items to. For each instance, we have a collection of rankings that is also given as input. The j-th ranking for the i-th instance is denoted $\sigma_j^{(i)}$. For example, in the case of our ensemble approach to classification, for each instance i we have a ranking of the labels given by the j-th input classifier, denoted $\sigma_j^{(i)} \in \mathcal{S}_n$. The permutation $\pi^{(i)} \in \mathcal{S}_n$ will be used to denote the predicted permutation for the i-th instance. We will denote a sequence of permutations σ_j as $\boldsymbol{\sigma}$, using boldface vector notation.

A function $d : \mathcal{S}_n \times \mathcal{S}_n \to \mathbb{R}$ is a *metric* on $\mathcal{S}_n$ if it satisfies the usual axioms:

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

For d to be a measure of distance between two rankings it also makes sense to require *invariance* of d over arbitrary relabeling of the items. This amounts to the following right invariance property:

$$d(\pi,\sigma) \;=\; d(\pi\tau,\sigma\tau) \quad \forall \pi,\sigma,\tau \in \mathcal{S}_n. \tag{1}$$

We list some possible choices for d that fulfill the metric and right invariance properties, and that make sense in the applications considered in this paper. These are Spearman's footrule F and rank correlation R, and Kendall's τ, referred to below as T. These metrics are defined by

$$R(\pi,\sigma) \;=\; \sum_{i=1}^{n} (\pi(i) - \sigma(i))^2 \tag{2}$$

$$F(\pi,\sigma) \;=\; \sum_{i=1}^{n} |\pi(i) - \sigma(i)| \tag{3}$$

$$T(\pi,\sigma) \;=\; \sum_{i=1}^{n-1} \sum_{l>i} I(\pi\sigma^{-1}(i) - \pi\sigma^{-1}(l)) \tag{4}$$

where $I(x) = 1$ for $x > 0$ and 0 otherwise. The measure $T(\pi,\sigma)$ can be interpreted as the minimum number of adjacent transpositions needed to bring π to σ. By an adjacent transposition we mean an operation that flips a pair of items that have adjacent ranks. Another distance that is simple to work with, but not as appropriate for the applications considered here, is the Cayley distance given by $S(\pi,\sigma) = n - C(\pi^{-1}\sigma)$, where $C(\eta)$ is the number of cycles in η; this is equivalent to the minimum number of *non-adjacent* transpositions needed to bring π to σ (Cayley, 1849). See Fligner and Verducci (1986) for an interpretation of T and S in terms of multistage ranking. The above three metrics can be transformed to the range $[-1, 1]$ by the mapping $d \mapsto 1 - 2d/M$, where $M = \max_{\pi,\sigma} d(\pi,\sigma)$.

Now, let $\mathcal{S}_{n-k}$ denote the subgroup of $\mathcal{S}_n$ consisting of all permutations that fix the first k positions:

$$\mathcal{S}_{n-k} \;=\; \{\pi \in \mathcal{S}_n \,|\, \pi(i) = i,\ \forall i = 1, \dots, k\}. \tag{5}$$

The right coset

$$\mathcal{S}_{n-k}\pi \;=\; \{\sigma\pi \,|\, \sigma \in \mathcal{S}_{n-k}\} \tag{6}$$

is equivalent to a partial ranking, where we only consider the k top-ranked items. The set of all partial rankings of k out of n elements forms the quotient group $\mathcal{S}_n/\mathcal{S}_{n-k}$.

A *partition of* n is a sequence $\gamma = n_1, \dots, n_r$ of positive integers that sum up to n. Such a partition corresponds to a partial ranking of n_1 items in first position, n_2 items in second position and so on. A partial ranking of the top k items is a special case with $r = k+1, n_1 = \ldots = n_k = 1, n_{k+1} = n - k$. Let $N_1 = \{1, \dots, n_1\}, N_2 = \{n_1 + 1, \dots, n_1 + n_2\}, \cdots, N_r = \{n_1 + \cdots + n_{r-1} + 1, \dots, n\}$. Then the subgroup $\mathcal{S}_\gamma = \mathcal{S}_{n_1} \times \cdots \times \mathcal{S}_{n_r}$ contains all permutations $\pi \in \mathcal{S}_n$ for which the set equality $\pi(N_i) = N_i$ holds for each i; that is, all permutations that only permute within N_i. A partial ranking of type γ is equivalent to a coset $\mathcal{S}_\gamma\pi$ and the set of such partial rankings forms the quotient group $\mathcal{S}_n/\mathcal{S}_\gamma$.

3. A Conditional Ranking Model

This section presents the conditional model that forms the basis of the new ensemble method.

3.1 Standard models

Our starting point is the Mallows model (Mallows, 1957). The parameters in this model are a pair $(\theta \in \mathbb{R}, \sigma \in \mathcal{S}_n)$; σ is the location parameter, and θ is a dispersion parameter. The model is given by the following exponential form:

$$p(\pi \mid \theta, \sigma) \quad = \quad e^{\theta\, d(\pi,\sigma) - \psi(\theta,\sigma)} \tag{7}$$

where $d(\cdot,\cdot)$ is a right invariant metric on the symmetric group, and ψ is the cumulant function, $\psi(\theta,\sigma) = \log \sum_{\pi \in \mathcal{S}_n} \exp(\theta\, d(\pi,\sigma))$. We denote this model by $\mathcal{M}_d(\theta,\sigma)$. When d is Spearman's rank correlation or Kendall's τ, then by the right invariance of the metric, the cumulant function is seen to be only a function of θ. That is, since

$$\sum_{\pi \in \mathcal{S}_n} e^{\theta d(\pi,\sigma)} \; = \; \sum_{\pi \in \mathcal{S}_n} e^{\theta d(\pi\sigma^{-1},1)} \; = \; \sum_{\pi' \in \mathcal{S}_n} e^{\theta d(\pi',1)} \tag{8}$$

the normalizing constant can be written as $Z(\theta) = e^{\psi(\theta)}$. The parameters of the model, θ and σ, are typically estimated by maximum likelihood. Note that for negative θ, the model becomes more concentrated around the mode σ as θ decreases.

Fligner and Verducci proposed an $n-1$ parameter generalization of the Mallows model that may be interpreted as multistage ranking (Fligner & Verducci, 1986). For Kendall's τ distance their model is given by

$$p(\pi \mid \boldsymbol{\theta}, \sigma) \quad = \quad \frac{1}{Z(\theta,\sigma)} e^{\sum_{i=1}^{n-1} \theta_i \sum_{l>i} I(\pi\sigma^{-1}(i) - \pi\sigma^{-1}(l))} \tag{9}$$

where now $\boldsymbol{\theta} \in \mathbb{R}^{n-1}$. This model reduces to the Mallows model with Kendall's τ when all the parameters θ_i are identical.

3.2 An extension to multiple input rankings

We propose a generalization of the Mallows model for estimating a conditional distribution, which is similar to the model suggested by Feigin in chapter 5 of (Fligner & Verducci, 1993). Let $\sigma_j \in \mathcal{S}_n$ be a permutation, and $\theta_j \in \mathbb{R}$ for $j = 1, \ldots k$. The distribution

$$p(\pi \mid \boldsymbol{\sigma}, \boldsymbol{\theta}) = \frac{1}{Z(\boldsymbol{\theta}, \boldsymbol{\sigma})} e^{\sum_{j=1}^{k} \theta_j\, d(\pi,\sigma_j)} \tag{10}$$

defines a conditional model when there are multiple instances and each instance is associated with a possibly different set of rankings $\sigma_j^{(i)}$. These may represent, for example, the ranking of Web pages output from individual search engines for a particular query, or the ordering of the class labels output by the classifiers in the ensemble for a particular instance. The parameters θ_j can be then thought of as indicating the "degree of expertise" of the different rankers. In this setting, only the θ_j are the free parameters to be estimated. As an exponential model, the likelihood function is convex in θ and enjoys many nice asymptotic properties.

When d is Kendall's τ, it is obvious how to extend this to a higher dimensional model in a manner that is analogous to the Fligner and Verducci model described above. While this may indeed be a very useful extension, we do not pursue it further in this paper.

3.3 A Bayesian interpretation

Although the above model is naturally viewed as a discriminative model for π, underlying it is a natural Bayesian interpretation. Viewing π as a parameter, now suppose that $\sigma_j \sim \mathcal{M}_d(\theta_j, \pi)$; that is, the σ_j are independently sampled from Mallows models with common mode π and dispersion parameters θ_j. Under a prior $p(\pi)$, the posterior is

$$p(\pi \mid \boldsymbol{\theta}, \boldsymbol{\sigma}) \;\propto\; p(\pi)\, \exp\left(\textstyle\sum_j \theta_j d(\pi, \sigma_j)\right) \tag{11}$$

since by invariance of $d(\cdot,\cdot)$, the normalizing constants in the Mallows models cancel. Thus, under a uniform prior on π, the distribution (10) is precisely the posterior of a generative model for the rankings $\boldsymbol{\sigma}$. See (Fligner & Verducci, 1990) for a discussion of posterior probabilities for multistage ranking models.

Both AdaBoost and additive logistic regression have models of the form $q(y \mid \boldsymbol{\theta}, x) \;\propto\; \exp\left(\sum_j \theta_j f_j(x,y)\right)$, and a decision rule that is a linear combination of features $\sum_j \theta_j f_j(x,y)$. While boosting and maximum likelihood logistic regression are intimately related, as shown in (Lebanon & Lafferty, 2001), our ranking model is fundamentally different, since it is a model over rankings π and not labels y. Furthermore, in the applications presented below, the features or weak learners are used to generate the rankings $\boldsymbol{\sigma}$, rather than to form an additive model. The above observation shows, however, that the model can be viewed as both a discriminative and (the posterior of) a generative model, a fact that may facilitate asymptotic analysis.

4. Learning and Inference

This section describes parameter estimation for the model presented in the previous section, as well as how it may be used for inference. These issues are non-standard here because the model is over rankings while the training and test data are often only annotated with individual class labels.

4.1 Learning

When supplied with a training set consisting of pairs $D = \left\{(\pi^{(i)}, \boldsymbol{\sigma}^{(i)})\right\}$, the parameters of the model can be estimated by maximum conditional likelihood or MAP. In principle, maximum likelihood estimation for this model is straightforward, and can be carried out using numerical algorithms such as a conjugate gradient procedure. However in practice, several obstacles may make training more challenging.

In many practical situations we do not have data in the form of full permutations. For example, in classification, for each training instance $x^{(i)}$, one is given the label $y^{(i)}$. In terms of a ranking model, this may be viewed as a partial permutation that corresponds to the coset $\mathcal{S}_{n-1}\sigma$ of the top class. Moreover, the classifiers for an instance $x^{(i)}$ may be a full or partial permutation. In the case of search engine fusion, each ranker may provide a list of, say, the top 10 documents, corresponding to a coset $\mathcal{S}_{n-10}\sigma$. Here the label will correspond to a coset $S_\gamma \pi^{(i)}$, whose form depends on the relevance annotations in the data.

By treating the available data as censored and the full ranking $\pi^{(i)}$ and $\sigma_j^{(i)}$ as the complete data the model (10) can be learned by maximizing the marginal conditional likelihood. Suppose that instance $x^{(i)}$ is given label $y^{(i)}$. Suppose that $\boldsymbol{\sigma}^{(i)}$ are complete rankings, given by the rankings of the labels on instance $x^{(i)}$, and that $\pi^{(i)}$ is an arbitrary permutation that ranks $y^{(i)}$ at the top. Then the log-likelihood function for the model is given by

$$\ell(\theta) = \log \prod_i \sum_{\pi \in \mathcal{S}_{n-1}\pi^{(i)}} p(\pi \,|\, \boldsymbol{\theta}, \boldsymbol{\sigma}^{(i)}) \tag{12}$$

$$= \sum_i \log \left(\frac{\sum_{\pi \in \mathcal{S}_{n-1}\pi^{(i)}} e^{\sum_j \theta_j d(\pi, \sigma_j^{(i)})}}{\sum_{\pi \in \mathcal{S}_n} e^{\sum_j \theta_j d(\pi, \sigma_j^{(i)})}} \right). \tag{13}$$

More generally, one may observe only a partial ranking for π corresponding to a coset $\mathcal{S}_\delta \pi^{(i)}$, and the input rankers themselves may only provide a partial ranking $\{S_{\gamma_j}\sigma_j^{(i)}\}$ for the labels of x. For example, in the search engine setting, $\mathcal{S}_{\gamma_j}\sigma_j^{(i)}$ is the partial ranking of pages by the j-th search engine for the i-th query and $\mathcal{S}_\delta \pi^{(i)}$ may be obtained as relevance feedback from the user.

This may again be treated as a censored data problem. Lacking additional information about the rankers, we assume that conditioned on the coset $\{S_{\gamma_j}\sigma_j\}$, the full ranking σ_j is uniform. Under this assumption, the marginal likelihood used to fit the model is given by

$$p(S_\delta \pi^{(i)} \,|\, \boldsymbol{\theta}, \{S_{\gamma_j}\sigma_j^{(i)}\}) = \sum_{\pi' \in S_\delta \pi^{(i)}} \frac{1}{\prod_j \left| S_{\gamma_j}\sigma_j^{(i)} \right|} \sum_{\{\sigma'_j \in S_{\gamma_j}\sigma_j^{(i)}\}} p(\pi' \,|\, \boldsymbol{\theta}, \boldsymbol{\sigma}'). \tag{14}$$

While this marginal looks especially unpleasant, the use of MCMC methods for this model is fairly straightforward, as explained next.

4.2 Using MCMC

Maximum likelihood estimation using a first order technique such as conjugate gradient requires the computation of the log-likelihood derivatives

$$\frac{\partial \ell(\boldsymbol{\theta})}{\partial \theta_j} = \sum_i \sum_{\pi \in \mathcal{S}_{n-1}\pi^{(i)}} d(\pi, \sigma_j^{(i)})\, p(\pi \,|\, \boldsymbol{\theta}, \boldsymbol{\sigma}^{(i)}, \mathcal{S}\pi^{(i)}) - \sum_i \sum_{\pi \in \mathcal{S}_n} d(\pi, \sigma_j^{(i)})\, p(\pi \,|\, \boldsymbol{\theta}, \boldsymbol{\sigma}^{(i)}). \tag{15}$$

In addition, the line search in conjugate gradient requires the evaluation of the log-likelihood, given in terms of an expectation by

$$\ell(\boldsymbol{\theta}) = \sum_i \log \left(E[1_{\mathcal{S}_{n-1}\pi^{(i)}} \,|\, \boldsymbol{\theta}, \boldsymbol{\sigma}^{(i)}] \right) \tag{16}$$

For small n, these sums over $n!$ items can be computed explicitly. For larger n, Markov chain Monte Carlo methods may be attractive.

Running an MCMC algorithm such as Metropolis-Hastings for the generalized Mallows model is relatively straightforward. A natural proposal distribution $q(\eta \,|\, \pi)$ is to move by random transpositions:

$$q(\eta \,|\, \pi) = \begin{cases} 1/\binom{n}{2} & \text{if } S(\eta, \pi) = 1 \\ 0 & \text{otherwise,} \end{cases} \tag{17}$$

where $S(\eta, \pi)$ is the Cayley distance. To sample from $p(\pi \,|\, \boldsymbol{\theta}, \boldsymbol{\sigma}^{(i)}, \mathcal{S}\pi^{(i)})$, we use a proposal distribution that simply fixes the elements given by $\pi^{(i)}$, and randomly transposes the remaining elements.

Diaconis and Hanlon (1992) analyze this MCMC algorithm in the special case of a Mallows model with the Cayley distance as the metric, showing that it is rapidly mixing. Unfortunately, the Cayley distance is not as appropriate for classification and meta-search applications, and no such result is known with Kendall's τ in its place, nor with the generalization of the Mallows model that is proposed in this paper. However, we have found MCMC to give reasonable results when sampling over $\mathcal{S}_n/\mathcal{S}_{n-k}$ for the meta-search experiments with a proposal based on adjacent transpositions.

4.3 Inference

At test time, partial rankings are again of interest. For example, in classification one may wish to compute the marginal probability that a given label y has rank one. In the experiments reported below, the extended Mallows model is evaluated using the *expected* rank of the true label, which is then used to order the items. For a given test instance $(x^{(i)}, y^{(i)})$, the expected rank is

$$E\left[\pi(y^{(i)}) \mid \boldsymbol{\theta}, \boldsymbol{\sigma}^{(i)}\right] = \sum_{k=1}^{n} k\, p(\pi(y) = k \mid \boldsymbol{\theta}, \boldsymbol{\sigma}^{(i)}) \quad (18)$$

$$= \sum_{k=1}^{n} k \sum_{\pi \in \mathcal{S}_{\delta_k} \pi_k^{(i)}} p(\pi \mid \boldsymbol{\theta}, \boldsymbol{\sigma}^{(i)})$$

where $\mathcal{S}_{\delta_k} \pi_k^{(i)}$ is the coset of permutations which fix $y^{(i)}$ in position k. The labels can then be ordered according to their expected ranks. When used together with probability of correctness (probability of having rank one), this provides a more meaningful measure than the standard error rate since it gives information on "how far off" the model is on a particular instance. In the search engine setting, one is interested in presenting a final ranked list to a user. This list can be formed by ranking the documents according to the expected ranking of documents.

Working with cosets through censoring offers an appealing and principled approach to multilabel classification problems. In multilabel classification, every instance may have several labels. The output of the features may be any coset. For example, the feature functions may assign confidence scores that yield either full or partial ranking of the classes, or they may be multilabel classifiers that output an arbitrary coset. The model (10) can then be used to calculate probabilities of cosets, corresponding to multilabel assignments.

5. Extensions

This section briefly discusses possible extensions to the above model that could lead to significantly more accurate rankers.

First, the Bayesian interpretation presented in Section 3 could be modified to use a non-uniform prior $p(\pi)$. When viewed in terms of the posterior, this term becomes equivalent to a "carrier" or default density $a(x)$ in an exponential family model $p(x) = a(x) \exp(\theta f(x) - \psi(\theta))$. Another straightforward extension involves adding feature interaction terms, analogous to those described in (Friedman et al., 2000). The resulting model would take the form

$$p(\pi \mid \boldsymbol{\theta}, \boldsymbol{\sigma}) \propto \exp\left(\sum_j \theta_j\, d(\pi, \sigma_j) + \sum_{kl} \theta_{kl}\, d(\pi, \sigma_k)\, d(\pi, \sigma_l)\right) . \quad (19)$$

Additional binary covariates $\nu_j(x)$ may be incorporated into the model as

$$p(\pi \mid \boldsymbol{\theta}, \boldsymbol{\sigma}, x) = $$

$$= \frac{1}{Z(\boldsymbol{\theta}, \boldsymbol{\sigma}, x)} \exp\left(\sum_{ij} \nu_i(x)\, \theta_{ij}\, d(\pi, \sigma_j)\right) \quad (20)$$

$$= \frac{1}{Z(\boldsymbol{\theta}, \boldsymbol{\sigma}, x)} \exp\left(\boldsymbol{\nu}(x)^\top \boldsymbol{\theta}\, d(\pi, \boldsymbol{\sigma})\right) \quad (21)$$

where now $\boldsymbol{\theta}$ is a matrix of parameters. Here θ_{ij} can be interpreted as the expertise of ranker j on instances x for which $\nu_i(x) = 1$. In the search engine setting, ν could be binary features of the query that suggest an area of high accuracy for one of the engines.

An alternative approach to censoring can be used to train a model directly with partial permutations. In particular, the metrics described in Section 2 can be extended to coset spaces in various ways, and the model could be defined directly on partial permutations. Critchlow (1980) describes ways of extending a metric $d : \mathcal{S}_n \times \mathcal{S}_n \to \mathbb{R}$ to a metric d^* on coset spaces $d^* : \mathcal{S}_n/\mathcal{S}_\gamma \times \mathcal{S}_n/\mathcal{S}_\delta \to \mathbb{R}$. For such a metric, the conditional model becomes $p(\mathcal{S}_\delta \pi \mid \{\mathcal{S}_{\gamma_j} \sigma_j\}, \boldsymbol{\theta}) \propto \exp\left(\sum_j \theta_j\, d^*(\mathcal{S}_\delta \pi, \mathcal{S}_{\gamma_j} \sigma_j)\right)$. The main advantage of such an extension is computational. Summing over the coset may be intractable, and the use of an extended metric reduces the computational effort significantly. However, the censored approach seems more motivated and intuitive than the extended metric approach.

6. Experimental Results

This section reports the results of experiments on four data sets. The first is a synthetic data set generated from a mixture of Gaussians in three dimensions. The next two are multiclass datasets from the UC Irvine repository, `Vehicle` and `Glass`. The last is a meta-search experiment using a collection of queries and relevance rankings from different search engines. The first three are classification datasets and the meta-search data has ranked lists of retrieved Web pages.

6.1 Classification experiments

The mixture of Gaussian dataset consists of 500 training and testing points sampled iid from five Gaussians, under a uniform class prior. In the `Vehicle` dataset four car models were photographed at different orientations, and from these images 18 numeric and geometric features were extracted from their silhouettes. The dataset includes approximately 1000 instances; the task is to predict the vehicle model from the measurements. The `Glass` dataset has 214 instances, each labeled with one of six classes. Each instance has 10 numeric attributes originating from chemical measurements, with the classes corresponding to different types of glass.

Because the goal is to combine diverse classifiers, we restricted each classifier to work only on one dimension of the input space (chosen randomly). In the first experiment probabilistic decision stumps were used as the weak learners. A probabilistic decision stump $t_{d,\eta}(y|x)$ is defined as

$$t_{d,\eta}(y'|x') = \begin{cases} |A|/|C| & \text{if } x'_d < \eta \\ |B|/|\overline{C}| & \text{otherwise} \end{cases} \quad (22)$$

	ϵ_{train}	ϵ_{test}	ρ_{train}	ρ_{test}
`Glass`				
Real AdaBoost.M2	0.53	0.59	1.94	2.15
Discrete AdaBoost.M2	0.56	0.60	2.17	2.32
Discrete AdaBoost.MH	0.55	0.61	2.24	2.46
Discrete Logloss	0.56	0.61	2.19	2.32
Real Logloss	0.53	0.59	1.94	2.16
Cranking	0.54	0.58	2.07	2.28
`Vehicle`				
Real AdaBoost.M2	0.56	0.63	1.97	2.15
Discrete AdaBoost.M2	0.69	0.70	2.37	2.40
Discrete AdaBoost.MH	0.68	0.70	2.34	2.39
Discrete Logloss	0.69	0.70	2.36	2.39
Real Logloss	0.55	0.62	1.96	2.13
Cranking	0.58	0.64	2.06	2.22

Table 1. Error rates and rank rates, comparing various forms of additive models. Each ensemble method used the same set of input classifiers, made up of a combination of random stumps, neural nets, and k-nearest neighbor classifiers.

where $C \subset D_{train}$ is the set of training examples that satisfy $x_d < \eta$, $A = \{x \in C : y(x) = y'\}$ and $B = \{x \in \overline{C} : y(x) = y'\}$.

In the first set of experiments, different methods for combining a set of randomly generated stumps were compared. Three general types of methods were used, depending on whether they used the confidence scores of the weak learners (real AdaBoost and real logistic regression), binary decisions only (discrete AdaBoost and discrete logistic regression), or the ranked list of labels output by the stump (cranking). When using binary or confidence scores, we formed a linear combination of the features by minimizing the exponential loss of AdaBoost.M2 or maximizing the likelihood in the case of logistic regression; the relationship between these optimization problems is discussed at length in (Friedman et al., 2000; Collins et al., 2002; Lebanon & Lafferty, 2001). When using the ranked data, the cranking model (10) was fit using a conjugate gradient algorithm to maximize the marginal likelihood. In each case, the parameters were fit using an iterative algorithm that makes multiple passes through the data.

To measure the performance of the ensembles we used the error rate ϵ and the rank rate ρ defined as the average of the ranks assigned to the correct labels. Error and rank rates over the train and test set for the five methods are plotted in the graphs in Figure 1. The plots were averaged using 10-fold cross validation, with the same train/test splits used for all methods. As expected, the use of confidence scores in the "real" versions of logistic regression and boosting yields better performance than the use of binary models. On the Gaussian mixture data, cranking significantly outperforms the other methods with respect to both the error rate ϵ_{test} and the rank rate ρ_{test}. For the `Vehicle` dataset, cranking outperforms the binary models but is worse than the methods that use confidence scores.

In an additional set of experiments, the same ensemble methods considered above were used for combining *different* classifiers; specifically, we combined randomly selected stumps, single layer feed-forward neural networks (using 1-of-n label encoding), and k-nearest neighbor classifiers. All three types of classifiers output confidence scores in the range $[0, 1]$. The results are displayed in Table 1.

6.2 Meta-search experiments

The second set of experiments were carried out using the meta-search dataset described in (Cohen et al., 1999). In this data, queries with the names of machine learning researchers and universities were first expanded in different ways, and then used to retrieve different lists of Web pages. The queries were designed in such a way that the correct Web page was known, making it possible to easily assign relevance judgements to the returned pages. (We used only the machine learning queries since for the university queries a single query expansion method outperforms the other engines for all queries.) After removing engines and queries that didn't have the necessary data, we were left with 14 rankers and 79 queries.

Recent work on the fusion problem includes RankBoost (Freund et al., 1998), the use of a linear combination of estimated relevance scores (Vogt & Cottrell, 1999), modeling the scoring process of the different search engines (Manmatha et al., 2001) and a naïve Bayes approach, referred to as Bayes Fuse in (Aslam & Montague, 2001). In the case of RankBoost, the hypothesis class is a linear combination of the ranks of the individual items assigned by the input rankers. In all of these approaches, a single ranking is output by the system rather than a distribution over rankings.

In the experiments presented below, we compare cranking with Bayes Fuse (BF). The BF model ranks documents according to the posterior log-odds score $\log(p(\mathit{rel} \,|\, r_1, \ldots, r_k)/p(\mathit{irr} \,|\, r_1, \ldots, r_k))$ where r_i is the rank assigned by engine i to the document in question. The log-odds were computed using Bayes rule and a naïve Bayes independence assumption

$$\log\left(\frac{p(\mathit{rel} \,|\, r_1, \ldots, r_k)}{p(\mathit{irr} \,|\, r_1, \ldots, r_k)}\right) = \log\left(\frac{p(\mathit{rel}) \prod_i p(r_i \,|\, \mathit{rel})}{p(\mathit{irr}) \prod_i p(r_i \,|\, \mathit{irr})}\right). \tag{23}$$

The probabilities $p(r_i \,|\, \mathit{rel})$ and $p(r_i \,|\, \mathit{irr})$ were computed using MAP estimation for a multinomial distribution with Laplace smoothing, as well as with a kernel estimate. Both models were used to rank the pages returned by all of the search engines.

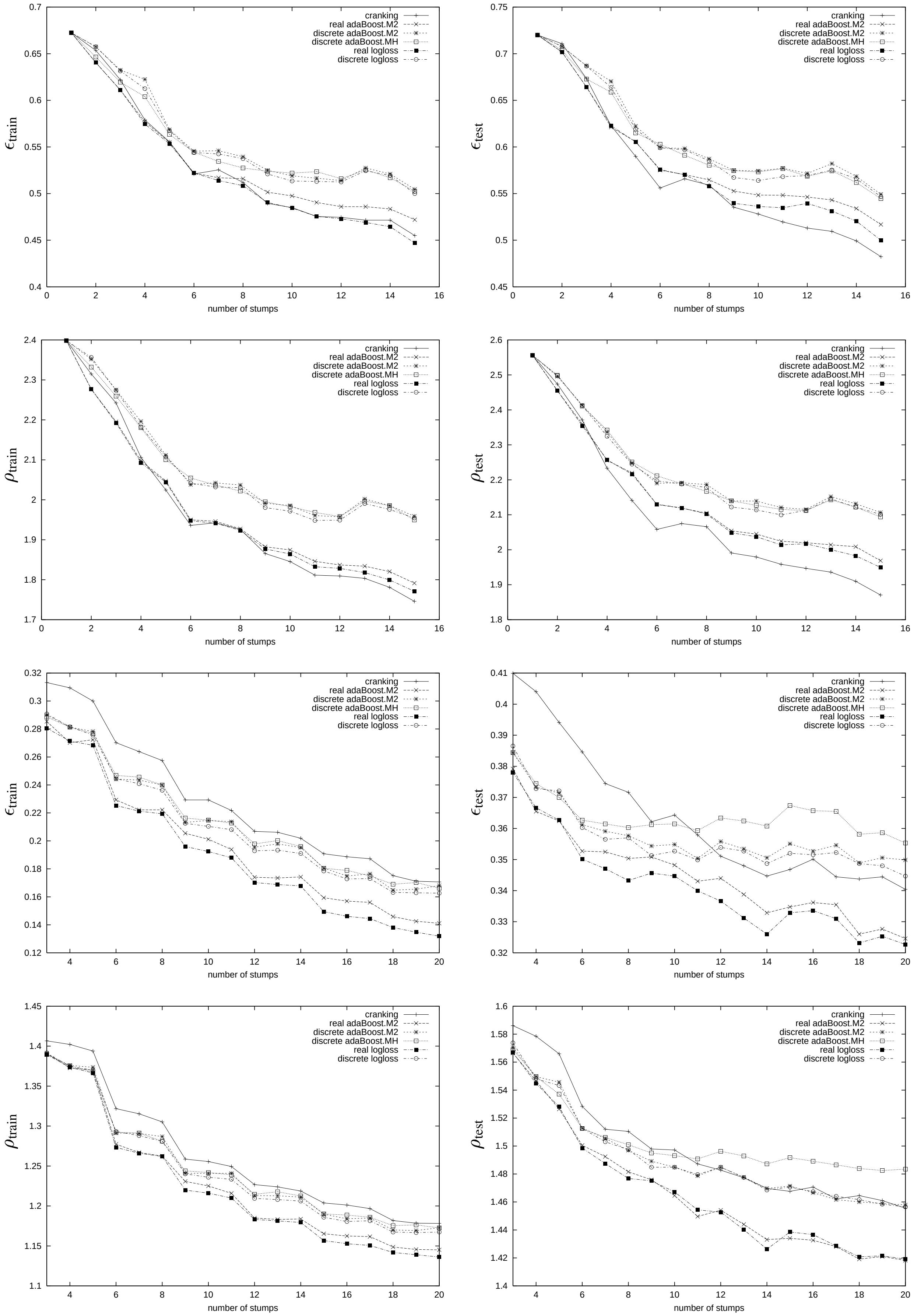

Figure 1. Error and rank rates for combining random stumps. Each plot tracks the error or rank rate as the number of features is increased. The same features are used in all methods. The first two rows are for the Gaussian mixture dataset and the last two rows are for `Vehicle`.

Coset size	ρ_{test}^{BF-L}	ρ_{test}^{BF-K}	ρ_{test}^{BEST-1}	ρ_{test}^{CM}
3	1.5	1.7	3.3	2.5
7	3.6	3.5	4.8	3.0
10	3.6	3.1	6.0	2.7

Table 2. Rank rate results for the meta-search data. When considering the top 7 or 10, the rank rate of cranking is lower than that of Bayes Fuse.

In each experiment the top l documents were considered from each engine, for $l = 3, 7$, or 10. The total number of pages n returned by all of the engines is large, however, so we use a model defined directly on the coset $\mathcal{S}_n/\mathcal{S}_{n-l}$, as described briefly in Section 5. Since there is a single relevant document for each query, fitting this model involves maximizing the marginal conditional likelihood. To do this, a gradient ascent algorithm was used, where the gradient and log-likelihood are estimated using a Metropolis-Hastings algorithm, as described in Section 4.

To evaluate the methods, we compute the average rank of the correct page in the test set queries, denoted $\rho_{\text{test}}^{BF-L}$ for the Bayes Fuse with Laplace smoothing and $\rho_{\text{test}}^{BF-K}$ for the Bayes Fuse kernel estimator. $\rho_{\text{test}}^{BEST-1}$ denotes the average rank rate for the best single ranker. Note that the best single ranker was found using both the train and test queries. The average rank for cranking, obtained by ordering the pages according to their expected rank with respect to (10) as described in Section 4.3, is denoted by ρ_{test}^{CM}. Note that all methods performed better than the best single ranker and as the number of documents retrieved by each ranker grows cranking outperforms Bayes Fuse.

7. Conclusions

This paper introduced a new approach to ensemble learning that takes ranking, rather than classification, as fundamental, leading to models on the symmetric group and its cosets. In this respect it is to the best of our knowledge unlike any previous work in the machine learning literature. An extension of the Mallows model was proposed as the basis for the new approach, which has both discriminative and generative interpretations due to the right invariance of the underlying metric. As a method for combining classifiers, the ranking approach has many attractive features; for example, it naturally models the assignment of multiple labels to instances. Experiments carried out on synthetic and UCI datasets for classification, as well as on meta-search data for ranking, indicate that the new approach compares well with strong competing methods. The framework draws on the significant body of research that has been carried out on ranking models in the statistical literature, comprising a rich theory that has significant promise for future development in machine learning.

Acknowledgments

We thank Joseph Verducci for helpful comments and William Cohen for providing the meta-search data.

References

Aslam, J. A., & Montague, M. (2001). Models for metasearch. *Proc. of the ACM SIGIR conference*.

Cayley, A. (1849). A note on the theory of permutations. *Phil. Mag.*, *34*, 527–529.

Cohen, W. C., Schapire, R. E., & Singer, Y. (1999). Learning to order things. *Journal of Artificial Intelligence Research*, *10*, 243–270.

Collins, M., Schapire, R. E., & Singer, Y. (2002). Logistic regression, AdaBoost and Bregman distances. *Machine Learning*, *48*, 253–285.

Critchlow, D. (1980). *Metric methods for analyzing partially ranked data*. Springer-Verlag.

Diaconis, P. (1988). *Group representations in probability and statistics*. Institute of Mathematical Statistics.

Diaconis, P. (1989). A generalization of spectral analysis with application to ranked data. *Annals of Statistics*, *17*, 949–979.

Diaconis, P., & Hanlon, P. (1992). Eigen analysis for some examples of the Metropolis algorithm. *Contemporary Mathematics*, *138*, 99–117.

Fligner, M. A., & Verducci, J. S. (1986). Distance based ranking models. *Journal of the Royal Statistical Society B*, *48*, 359–369.

Fligner, M. A., & Verducci, J. S. (1988). Multistage ranking models. *Journal of the American Statistical Association*, *88*, 892–901.

Fligner, M. A., & Verducci, J. S. (1990). Posterior probabilities for a concensus ordering. *Psychometrika*, *55*, 53–63.

Fligner, M. A., & Verducci, J. S. (Eds.). (1993). *Probability models and statistical analyses for ranking data*. Springer-Verlag.

Freund, Y., Iyer, R., Schapire, R. E., & Singer, Y. (1998). An efficient boosting algorithm for combining preferences. *International Conference on Machine Learning*.

Freund, Y., & Schapire, R. E. (1996). Experiments with a new boosting algorithm. *International Conference on Machine Learning*.

Friedman, J., Hastie, T., & Tibshirani, R. (2000). Additive logistic regression: A statistical view of boosting. *The Annals of Statistics*, *28*, 337–407 (with discussion).

Lebanon, G., & Lafferty, J. (2001). Boosting and maximum likelihood for exponential models. *Neural Information Processing Systems (NIPS)*.

Mallows, C. L. (1957). Non-null ranking models. *Biometrika*, *44*, 114–130.

Manmatha, R., Rath, T., & Feng, F. (2001). Modeling score distributions for combining the outputs of search engines. *Proc. of the 24th ACM SIGIR conference*.

Smith, B. B. (1950). Discussion of professor Ross's paper. *Journal of the Royal Statistical Society B*, *12*, 53–56.

Thurstone, L. L. (1927). A law of comparative judgement. *Psychological Review*, *34*, 273–286.

Vogt, C. C., & Cottrell, G. W. (1999). Fusion via a linear combination of scores. *Information Retrieval Journal*, *1*, 151–173.