# Predicting Partial Orders: Ranking with Abstention

Weiwei Cheng[1], Michaël Rademaker[2], Bernard De Baets[2], and Eyke Hüllermeier[1]

[1] Department of Mathematics and Computer Science
University of Marburg, Germany
{cheng,eyke}@mathematik.uni-marburg.de
[2] Department of Applied Mathematics, Biometrics and Process Control
Ghent University, Belgium
{michael.rademaker,bernard.debaets}@ugent.be

**Abstract.** The prediction of structured outputs in general and rankings in particular has attracted considerable attention in machine learning in recent years, and different types of ranking problems have already been studied. In this paper, we propose a generalization or, say, relaxation of the standard setting, allowing a model to make predictions in the form of partial instead of total orders. We interpret such kind of prediction as a ranking with partial abstention: If the model is not sufficiently certain regarding the relative order of two alternatives and, therefore, cannot reliably decide whether the former should precede the latter or the other way around, it may abstain from this decision and instead declare these alternatives as being incomparable. We propose a general approach to ranking with partial abstention as well as evaluation metrics for measuring the correctness and completeness of predictions. For two types of ranking problems, we show experimentally that this approach is able to achieve a reasonable trade-off between these two criteria.

## 1 Introduction

The problem of "learning to rank" has recently attracted considerable attention in machine learning, and different types of ranking problems have been studied, both theoretically and empirically. Roughly speaking, the goal of methods developed in this field is to learn a "ranker" that outputs predictions in the form of a ranking of a set of alternatives. Thus, learning to rank can be seen as a specific type of structured output prediction [1].

A ranking is commonly understood as a strict total order, i.e., an irreflexive, asymmetric, and transitive relation. In this paper, we propose a generalization of the standard setting, allowing a model to make predictions in the form of *partial* instead of *total* orders. We interpret such kind of prediction as a ranking with partial abstention: If the ranker is not sufficiently certain regarding the relative order of two alternatives and, therefore, cannot reliably decide whether the former should precede the latter or the other way around, it may abstain from this decision and instead declare these alternatives as being incomparable.

The notion of abstention is actually well-known for conventional classification, and the corresponding extension is usually referred to as *classification with a reject option* [2–4]: The classifier is allowed to abstain from a prediction for a query instance in case it is not sure enough. An abstention of this kind is an obvious means to avoid unreliable predictions. Needless to say, the same idea does also make sense in the context of ranking. In fact, one may even argue that a reject option becomes even more interesting here: While a conventional classifier has only two choices, namely to predict a class or to abstain, a ranker can abstain *to a certain degree*: The order relation predicted by the ranker can be more or less complete or, stated differently, more or less partial, ranging from a total order (conventional ranking) to the empty relation in which all alternatives are incomparable. Later on, we will express the degree of abstention of a ranker more precisely in terms of a degree of completeness of the partial order it predicts.

The main contribution of this paper is a general approach to ranking with partial abstention, which is applicable to different types of ranking problems. In a nutshell, our approach consists of two main steps. First, a preference relation is derived that specifies, for each pair of alternatives $\boldsymbol{a}$ and $\boldsymbol{b}$, a degree of preference for $\boldsymbol{a}$ over $\boldsymbol{b}$ and, vice versa, a degree of preference for $\boldsymbol{b}$ over $\boldsymbol{a}$. The idea is that, the more similar these two degrees are, the more uncertain the learner is. Then, in a second step, a partial order maximally compatible with this preference relation, in a sense to be specified later on, is derived as a prediction. In order to realize the first step, we make use of ensemble learning techniques, although other possibilities are conceivable.

The remainder of the paper is organized as follows. In the next section, we briefly review some important ranking problems. Our approach to ranking with partial abstention is then detailed in Section 3. In Section 4, we address the question of how to evaluate predictions in the form of partial orders and propose suitable performance metrics for measuring the correctness and completeness of such predictions. Section 5 is devoted to experimental studies. For two types of ranking problems, we show that our approach is indeed able to achieve a reasonable trade-off between these two criteria. The paper ends with a couple of concluding remarks in Section 6.

## 2 Ranking Problems

Following [5], we distinguish three types of ranking problems that have been studied extensively in the machine learning literature, namely label ranking [6–8], instance ranking [9], and object ranking [10], to be described in more detail in the following.

### 2.1 Label Ranking

Like in the conventional setting of supervised learning (classification), we assume to be given an instance space $\mathbf{X}$ and a finite set of labels $\mathbf{Y} = \{y_1, y_2, \ldots, y_k\}$.

In label ranking, the goal is to learn a "label ranker" in the form of an $\mathbf{X} \to S_{\mathbf{Y}}$ mapping, where the output space $S_{\mathbf{Y}}$ is given by the set of all total orders (permutations) of the set of labels $\mathbf{Y}$ (the notation is leaned on the common notation $S_k$ for the symmetric group of order $k$). Thus, label ranking can be seen as a generalization of conventional classification, where a complete ranking

$$y_{\pi_{\boldsymbol{x}}^{-1}(1)} \succ_{\boldsymbol{x}} y_{\pi_{\boldsymbol{x}}^{-1}(2)} \succ_{\boldsymbol{x}} \ldots \succ_{\boldsymbol{x}} y_{\pi_{\boldsymbol{x}}^{-1}(k)}$$

is associated with an instance $\boldsymbol{x}$ instead of only a single class label. Here, $\pi_{\boldsymbol{x}}$ is a permutation of $\{1, 2, \ldots, k\}$ such that $\pi_{\boldsymbol{x}}(i)$ is the position of label $y_i$ in the ranking associated with $\boldsymbol{x}$.

The training data $\mathcal{T}$ used to induce a label ranker typically consists of a set of pairwise preferences of the form $y_i \succ_{\boldsymbol{x}} y_j$, suggesting that, for instance $\boldsymbol{x}$, $y_i$ is preferred to $y_j$. In other words, a single "observation" consists of an instance $\boldsymbol{x}$ together with an ordered pair of labels $(y_i, y_j)$.

To measure the predictive performance of a label ranker, a loss function on rankings is needed. In principle, any distance or correlation measure on rankings (permutations) can be used for that purpose. An important example is Kendall's tau, which counts the number of pairs of labels that are incorrectly ordered and normalizes this number to the interval $[-1, +1]$: For two permutations $\pi$ and $\sigma$, let $c$ be the number of correctly ordered pairs $(i, j) \in \{1, \ldots, k\}^2$, i.e., the pairs $(i, j)$ with $i < j$ and $(\pi(i) - \pi(j))(\sigma(i) - \sigma(j)) > 0$. Likewise, let $d$ be the number of incorrectly ordered pairs, i.e., the pairs $(i, j)$ with $(\pi(i) - \pi(j))(\sigma(i) - \sigma(j)) < 0$. Kendall's tau, expressing a degree of correlation between $\pi$ and $\sigma$, is then given by

$$\tau = \frac{c - d}{k(k-1)/2} \tag{1}$$

This coefficient assumes the extreme value 1 if $\sigma = \pi$ and the value $-1$ if $\sigma$ is the reversal of $\pi$.

### 2.2 Instance Ranking

This setting proceeds from the setting of *ordinal classification*, where an instance $\boldsymbol{x} \in \mathbf{X}$ belongs to one among a finite set of classes $\mathbf{Y} = \{y_1, y_2, \ldots, y_k\}$ and, moreover, the classes have a natural order: $y_1 < y_2 < \ldots < y_k$. Training data consists of a set $\mathcal{T}$ of labeled instances. As an example, consider the assignment of submitted papers to categories *reject*, *weak reject*, *weak accept*, and *accept*.

In contrast to conventional classification, the goal is not to learn a classifier but a ranking function $f(\cdot)$. Given a subset $X \subset \mathbf{X}$ of instances as an input, the function produces a ranking, i.e., a (strict) total order $\succ$, of these instances as an output (typically by assigning a score to each instance and then sorting by scores).

For the case $k = 2$, this problem is well-known as the *bipartite ranking* problem. The case $k > 2$ has recently been termed *multipartite ranking* [9]. As an example, consider the task of a reviewer who has to rank the papers according to

their quality, possibly though not necessarily with the goal of partitioning this ranking into the above four categories.

Thus, the goal of *instance ranking* is to produce a ranking $\succ$ in which instances from higher classes precede those from lower classes. Different types of accuracy measures have been proposed for predictions of this kind. Typically, they count the number of ranking errors, that is, the number of pairs $(\boldsymbol{x}, \boldsymbol{x}') \in X \times X$ such that $\boldsymbol{x}$ is ranked higher than $\boldsymbol{x}'$ even though the former belongs to a lower class than the latter. In the two-class case, this amounts to the well-known AUC, the area under the ROC-curve [11]:

$$\mathrm{AUC}(\succ, X) = \frac{1}{|P| \cdot |N|} \sum_{\boldsymbol{x} \in P} \sum_{\boldsymbol{x}' \in N} \begin{cases} 1 & \text{if } \boldsymbol{x} \succ \boldsymbol{x}' \\ 0 & \text{if } \boldsymbol{x}' \succ \boldsymbol{x} \end{cases} , \tag{2}$$

where $P \subset X$ is the set of positive and $N \subset X$ the set of negative examples in $X$.[3] Its generalization to multiple (ordered) classes is known as the concordance index or C-index in statistics [12].

### 2.3 Object Ranking

In the setting of object ranking, there is no supervision in the sense that no output or class label is associated with an object. The goal in object ranking is to learn a ranking function $f(\cdot)$ which, given a subset $Z$ of an underlying referential set $\mathbf{Z}$ of objects as an input, produces a ranking of these objects as an output. Again, this is typically done by assigning a score to each instance and then sorting by scores.

Objects $\boldsymbol{z} \in \mathbf{Z}$ are commonly though not necessarily described in terms of an attribute-value representation. As training information, an object ranker has access to exemplary rankings or pairwise preferences of the form $\boldsymbol{z} \succ \boldsymbol{z}'$ suggesting that $\boldsymbol{z}$ should be ranked higher than $\boldsymbol{z}'$. This scenario is also known as "learning to order things" [10].

The performance of an object ranker can again be measured in terms of a distance function or correlation measure on rankings. In contrast to the setting of label ranking, however, the number of items to be ordered in the context of object ranking is typically much larger. Therefore, one often prefers measures that put more emphasis on the top of a ranking while paying less attention to the bottom [13]. In Web search, for example, people normally look at the top-10 results while ignoring the rest. Besides, the target is often not a "true" ranking but instead a single object or a subset of relevant objects, for example a set of documents relevant to a query. Evaluation measures especially tailored toward these types of requirements have been proposed in information retrieval. Typical examples include precision and recall as well as normalized discounted cumulative gain (NDCG) [14].

[3] Note that we assume $\succ$ to be a strict order. If ties are allowed, then these are typically counted by 1/2.

## 3 Ranking with Partial Abstention

As explained above, the set of alternatives, say, $\mathbf{A}$, to be ordered by a ranker depends on the type of ranking problem. In label ranking, $\mathbf{A}$ is a fixed set of labels $\mathbf{Y}$, whereas in instance ranking, it is a subset $X$ of the instance space $\mathbf{X}$. A ranking on $\mathbf{A}$ is a strict, total, asymmetric, and transitive relation $\succ$, specifying for all pairs $\boldsymbol{a}, \boldsymbol{b} \in \mathbf{A}$ whether $\boldsymbol{a}$ precedes $\boldsymbol{b}$, denoted $\boldsymbol{a} \succ \boldsymbol{b}$, or $\boldsymbol{b}$ precedes $\boldsymbol{a}$. The key property of transitivity can be seen as a principle of consistency: If $\boldsymbol{a}$ is preferred to $\boldsymbol{b}$ and $\boldsymbol{b}$ is preferred to $\boldsymbol{c}$, then $\boldsymbol{a}$ must be preferred to $\boldsymbol{c}$.

A partial order $\sqsupset$ on $\mathbf{A}$ is a generalization that sticks to this consistency principle but is not necessarily total. If, for two alternatives $\boldsymbol{a}$ and $\boldsymbol{b}$, neither $\boldsymbol{a} \sqsupset \boldsymbol{b}$ nor $\boldsymbol{b} \sqsupset \boldsymbol{a}$, then these alternatives are considered as incomparable, written $\boldsymbol{a} \perp \boldsymbol{b}$. Note that, in the following, we still assume strictness of $\sqsupset$, even of this is not always mentioned explicitly.

### 3.1 Partial Orders in Learning to Rank

As mentioned before, our idea is to make use of the concept of a partial order in a machine learning context, namely to generalize the problem of learning to rank. More specifically, the idea is that, for each pair of alternatives $\boldsymbol{a}$ and $\boldsymbol{b}$, the ranker can decide whether to make a prediction about the order relation between these labels, namely to hypothesize that $\boldsymbol{a}$ precedes $\boldsymbol{b}$ or that $\boldsymbol{b}$ precedes $\boldsymbol{a}$, or to abstain from this prediction. We call a ranker having this possibility of abstention a ranker with partial reject option. Note, however, that for different pairs of alternatives, the reject decisions cannot be made independently of each other. Instead, the pairwise predictions should of course be consistent in the sense of being transitive and acyclic. In other words, a ranker with a (partial) reject option is expected to make a prediction in the form of a (strict) partial order $\sqsupset$ on the set of alternatives. This partial order is considered as an incomplete estimation of an underlying (ground-truth) order relation $\succ$: For alternatives $\boldsymbol{a}, \boldsymbol{b} \in \mathbf{A}$, $\boldsymbol{a} \sqsupset \boldsymbol{b}$ corresponds to the prediction that $\boldsymbol{a} \succ \boldsymbol{b}$ (and not $\boldsymbol{b} \succ \boldsymbol{a}$) holds, whereas $\boldsymbol{a} \perp \boldsymbol{b}$ indicates an abstention on this pair of alternatives.

In this section, we propose a method that enables a ranker to make predictions of such kind. Roughly speaking, our approach consists of two main steps, to be detailed in the forthcoming subsections:

- The first step is the prediction of a preference relation $P$ that specifies, for each pair of alternatives $\boldsymbol{a}$ and $\boldsymbol{b}$, a degree of uncertainty regarding their relative comparison.
- In the second step, a (strict) partial order maximally compatible with this preference relation is derived.

### 3.2 Prediction of a Binary Preference Relation

Let $P$ be an $\mathbf{A} \times \mathbf{A} \to [0, 1]$ mapping, so that $P(\boldsymbol{a}, \boldsymbol{b})$ is a measure of support for the order (preference) relation $\boldsymbol{a} \succ \boldsymbol{b}$. We assume $P$ to be reciprocal, i.e.,

$$P(\boldsymbol{b}, \boldsymbol{a}) \;=\; 1 - P(\boldsymbol{a}, \boldsymbol{b})$$

for all $\boldsymbol{a}, \boldsymbol{b} \in \mathbf{A}$. A relation of that kind can be produced in different ways. For example, some ranking methods explicitly train models that compare alternatives in a pairwise way, e.g., by training a single classifier for each pair of alternatives [15]. If these models are able to make probabilistic predictions, these can be used directly as preference degrees $P(\boldsymbol{a}, \boldsymbol{b})$.

However, since probability estimation is known to be a difficult problem, we like to emphasize that our method for predicting strict partial orders does only assume an *ordinal* structure of the relation $P$. In fact, as will be seen below, the partial order induced by $P$ is invariant toward monotone transformations of $P$. In other words, only the order relation of preference degrees is important, not the degrees themselves: If $P(\boldsymbol{a}, \boldsymbol{b}) > P(\boldsymbol{a}', \boldsymbol{b}')$, then $\boldsymbol{a} \succ \boldsymbol{b}$ is considered as more certain than $\boldsymbol{a}' \succ \boldsymbol{b}'$.

Here, we propose a generic approach that allows one to turn every ranker into a partial ranker. To this end, we resort to the idea of ensembling. Let $L$ be a learning algorithm that, given a set of training data, induces a model $M$ that in turn makes predictions in the form of rankings (total orders) $\succ$ of a set of alternatives $\mathbf{A}$. Now, instead of training a single model, our idea is to train $k$ such models $M_1, \ldots, M_k$ by resampling from the original data set, i.e., by creating $k$ bootstrap samples and giving them as input to $L$. Consequently, by querying all these models, $k$ rankings $\succ_1, \ldots, \succ_k$ will be produced instead of a single prediction.

For each pair of alternatives $\boldsymbol{a}$ and $\boldsymbol{b}$, we then define the degree of preference $P(\boldsymbol{a}, \boldsymbol{b})$ in terms of the fraction of rankings in which $\boldsymbol{a}$ precedes $\boldsymbol{b}$:

$$P(\boldsymbol{a}, \boldsymbol{b}) \;=\; \frac{1}{k} \left| \{ i \mid \boldsymbol{a} \succ_i \boldsymbol{b} \} \right| \tag{3}$$

Thus, $P(\boldsymbol{a}, \boldsymbol{b}) = 1$ suggests a consensus among the ensemble members, since all of them agree that $\boldsymbol{a}$ should precede $\boldsymbol{b}$. On the other hand, $P(\boldsymbol{a}, \boldsymbol{b}) \approx 1/2$ indicates a highly uncertain situation.

### 3.3 Prediction of a Strict Partial Order Relation

On the basis of the preference relation $P$, we seek to induce a (partial) order relation $\sqsupset$ on $\mathbf{A}$, that we shall subsequently also denote by $\mathcal{R}$. Thus, $\mathcal{R}$ is an $\mathbf{A} \times \mathbf{A} \to \{0, 1\}$ mapping or, equivalently, a subset of $\mathbf{A} \times \mathbf{A}$, where $\mathcal{R}(\boldsymbol{a}, \boldsymbol{b}) = 1$, also written as $(\boldsymbol{a}, \boldsymbol{b}) \in \mathcal{R}$ or $\boldsymbol{a}\,\mathcal{R}\,\boldsymbol{b}$, indicates that $\boldsymbol{a} \sqsupset \boldsymbol{b}$.

The simplest idea is to let $\boldsymbol{a}\,\mathcal{R}\,\boldsymbol{b}$ iff $P(\boldsymbol{a}, \boldsymbol{b}) = 1$. The relation $\mathcal{R}$ thus defined is indeed a (strict) partial order, but since a perfect consensus ($P(\boldsymbol{a}, \boldsymbol{b}) \in \{0, 1\}$) is a strong requirement, most alternatives will be declared incomparable. Seeking a prediction that is as informative as possible, it is therefore natural to reduce the required degree of consensus. We therefore proceed from an "$\alpha$-cut" of the relation $P$, defined as

$$\mathcal{R}_\alpha = \{ (\boldsymbol{a}, \boldsymbol{b}) \mid P(\boldsymbol{a}, \boldsymbol{b}) \geq \alpha \} \tag{4}$$

for $0 < \alpha \leq 1$. A cut of that kind provides a reasonable point of departure, as it comprises the most certain preference statements while ignoring those comparisons $(\boldsymbol{a}, \boldsymbol{b})$ with $P(\boldsymbol{a}, \boldsymbol{b}) < \alpha$. However, it is not necessarily transitive and may even contain cycles. For example, suppose $\boldsymbol{a} \succ_1 \boldsymbol{b} \succ_1 \boldsymbol{c}$, $\boldsymbol{b} \succ_2 \boldsymbol{c} \succ_2 \boldsymbol{a}$ and $\boldsymbol{c} \succ_3 \boldsymbol{a} \succ_3 \boldsymbol{b}$. Clearly, $P(\boldsymbol{a}, \boldsymbol{b}) = P(\boldsymbol{b}, \boldsymbol{c}) = P(\boldsymbol{c}, \boldsymbol{a}) = 2/3$, rendering $\mathcal{R}_{2/3}$ a cyclical relation. While transitivity is easily enforced by computing the transitive closure of $\mathcal{R}_\alpha$, absence of cycles is not as easily obtained. Intuitively, it seems natural that for larger $\alpha$, cycles become less probable. However, as the example shows, even for $\alpha > 1/2$, cycles can still occur. Furthermore, the larger $\alpha$, the less informative the corresponding $\mathcal{R}_\alpha$.

Consequently, we propose to look for a minimal $\alpha$ (denote it as $\alpha^*$) such that the transitive closure of $\mathcal{R}_\alpha$ (denote it as $\overline{\mathcal{R}}_\alpha$) is a strict partial order relation [16]. This $\overline{\mathcal{R}}_{\alpha^*}$ will be the predicted strict partial order relation $\mathcal{R}$, and we call $\alpha^*$ the consensus threshold. By minimizing this threshold, we maximize $\mathcal{R}_\alpha$ as well as its transitive closure $\overline{\mathcal{R}}_\alpha$, and thereby also the information extracted from the ensemble on the basis of which $P$ was computed. In the remainder of this section, we deal with the problem of computing $\alpha^*$ in an efficient way.

### 3.4 Determination of an Optimal Threshold

Suppose that $P$ can assume only a finite number of values. In our case, according to (3), this set is given by $D = \{0, 1/k, 2/k, \ldots, 1\}$, and its cardinality by $k+1$, where $k$ is the ensemble size. Obviously, the domain of $\alpha$ can then be restricted to $D$. The simplest approach, therefore, it to test each value in $D$, i.e., to check for each value whether $\mathcal{R}_\alpha$ is acyclic, and hence $\overline{\mathcal{R}}_\alpha$ a partial order. Of course, instead of trying all values successively, it makes sense to exploit a monotonicity property: If $\mathcal{R}_\alpha$ is not acyclic, then $\mathcal{R}_\beta$ cannot be acyclic either, unless $\beta > \alpha$. Consequently, $\alpha^*$ can be found in at most $\log_2(k+1)$ steps using bisection. More specifically, by noting that $\alpha^*$ is lower-bounded by

$$\alpha_l = \frac{1}{k} + \max_{\boldsymbol{a},\boldsymbol{b}} \min\left(P(\boldsymbol{a}, \boldsymbol{b}), P(\boldsymbol{b}, \boldsymbol{a})\right) \tag{5}$$

and trivially upper-bounded by $\alpha_u = 1$, one can repeatedly update the bounds as follows, until $\alpha_u - \alpha_l < 1/k$:

(i) set $\alpha$ to the middle point between $\alpha_l$ and $\alpha_u$
(ii) compute $\mathcal{R}_\alpha$
(iii) compute $\overline{\mathcal{R}}_\alpha$ (e.g., using the Floyd-Warshall's algorithm [17])
(iv) if $\overline{\mathcal{R}}_\alpha$ is a partial order, set $\alpha_u$ to $\alpha$
(v) else set $\alpha_l$ to $\alpha$

This procedure stops with $\alpha^* = \alpha_l$. The complexity of this procedure is not worse than the transitive closure operation, i.e., it is at most $\mathcal{O}(|\mathbf{A}|^3)$.

As shown in [16], the same result can be computed with another algorithm that is conceptually simpler (though equally costly in terms of complexity, at least theoretically). This algorithm operates on an $|\mathbf{A}| \times |\mathbf{A}|$ matrix $\mathbf{R}$ initialized

with the entries $P(\boldsymbol{a}, \boldsymbol{b})$ (recall that $\mathbf{A}$ is the set of alternatives). It repeatedly performs a transitive closure operation at all the levels of $D$ simultaneously:

$$\mathbf{R}(\boldsymbol{a}, \boldsymbol{b}) \leftarrow \max\left(\mathbf{R}(\boldsymbol{a}, \boldsymbol{b}), \max_{\boldsymbol{c} \in \mathbf{A}}(\min(\mathbf{R}(\boldsymbol{a}, \boldsymbol{c}), \mathbf{R}(\boldsymbol{c}, \boldsymbol{b}))\,)\right) \tag{6}$$

for all $\boldsymbol{a}, \boldsymbol{b} \in \mathbf{A}$, until no further changes occur. These transitive closure operations can be seen as a correction of inconsistences in $P$ ($\boldsymbol{a}$ is to some degree preferred to $\boldsymbol{b}$, which in turn is to some degree preferred to $\boldsymbol{c}$, but $\boldsymbol{a}$ is not sufficiently preferred to $\boldsymbol{c}$). Since these inconsistencies do not occur very often, the number of update operations needed to stabilize $\mathbf{R}$ is normally quite small; in practice, we found that we rarely need more then one or two iterations.

**Algorithm 1**

**Require:** training data $\mathcal{T}$, test data $\mathcal{D}$, ensemble size $k$, base learner $L$
**Ensure:** a matrix $\mathbf{R}$ encoding partial order information for alternatives in $\mathcal{D}$ ($\mathbf{R}(i,j) = 1$ means $d_i \succ d_j$, where $d_i, d_j \in \mathcal{D}$)

initialize $\mathbf{R}$ as zero matrix
generate $k$ bootstrap samples from $\mathcal{T}$
constitute the ensemble with $k$ rankers trained using $L$
get $k$ rankings of alternatives in $\mathcal{D}$
**for** each of $k$ rankings **do**
  **for** every pair of alternatives $d_i, d_j \in \mathcal{D}$ **do**
    **if** $d_i \succ d_j$ **then**
      set $\mathbf{R}(i,j) := \mathbf{R}(i,j) + 1/k$
    **end if**
  **end for**
**end for**
**repeat**
  **for** every entry in $\mathbf{R}$ **do**
    $\mathbf{R}(i,j) := \max\left(\mathbf{R}(i,j), \max_{k \in \mathcal{D}}(\min(\mathbf{R}(i,k), \mathbf{R}(k,j)))\right)$
  **end for**
**until** No entry in $\mathbf{R}$ is changed.
**for** every entry in $\mathbf{R}$ **do**
  $\alpha := \max_{i,j} \min(\mathbf{R}(i,j), \mathbf{R}(j,i))$
**end for**
**for** every entry in $\mathbf{R}$ **do**
  **if** $\mathbf{R}(i,j) > \alpha$ **then**
    $\mathbf{R}(i,j) := 1$
  **end if**
**end for**

By construction, thresholding the final relation $\mathbf{R}$ at a level $\alpha$ will yield the transitive closure of relation $\mathcal{R}_\alpha$ in (4). Therefore, $\alpha^*$ can be taken as

$$\alpha^* = \frac{1}{k} + \max\left(\mathbf{R}(\boldsymbol{a}, \boldsymbol{b}) \,|\, \mathbf{R}(\boldsymbol{a}, \boldsymbol{b}) \leq \mathbf{R}(\boldsymbol{b}, \boldsymbol{a})\right) , \tag{7}$$

which is obviously the smallest $\alpha$ that avoids cycles. The whole procedure is summarized in Algorithm 1.

Finally, we note that, as postulated above, $\alpha^*$ in (7) yields a maximal partial order as a prediction. In principle, of course, any larger value can be used as well, producing a less complete relation and, therefore, a more "cautious" prediction. We shall come back to this issue in Section 4.

### 3.5 Illustrating Example

We illustrate our approach by means of a small two-dimensional toy example for the case of bipartite ranking. Suppose that the conditional class distributions of the positive and the negative class are two overlapping Gaussians. A training data set may then look like the one depicted in Fig. 1 (left), with positive examples as black and negative examples as white dots. Given a new set of query instances $X$ to be ranked, one may expect that a learner will be uncertain for those instances lying close to the overlap region, and may hence prefer to abstain from comparing them.

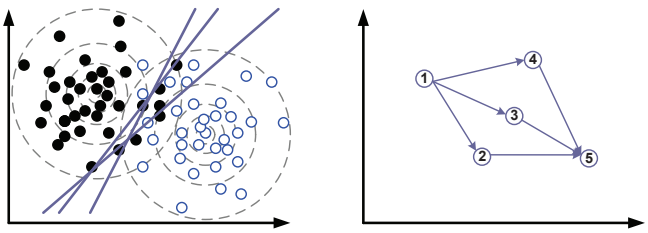



**Fig. 1.** Left: training data and ensemble models; right: partial order predicted for a set of five query instances.

More specifically, suppose that a linear model is used to train a ranker. Roughly speaking, this means fitting a separating line and sorting instances according to their distance from the decision boundary. Fig. 1 (left) shows several such models that may result from different bootstrap samples. Now, consider the five query instances shown in the right picture of Fig. 1. Whereas all these models will rank instance 1 ahead of 2, 3 and 4, and these in turn ahead of 5, instances 2, 3 and 4 will be put in various orders. Applying our approach as outlined above, with a proper choice of the threshold $\alpha$, may then yield the strict partial order indicated by the arrows in the right picture of Fig. 1. A prediction of that kind agrees with our expectation: Instance 1 is ranked first and instance 5 last; instance 2, 3 and 4 are put in the middle, but the learner abstains from comparing them in a mutual way.

## 4 Evaluation Measures

If a model is allowed to abstain from making predictions, it is expected to reduce its error rate. In fact, it can trivially do so, namely by rejecting all predictions, in which case it avoids any mistake. Clearly, this is not a desirable solution. Indeed, in the setting of prediction with reject option, there is always a trade-off between two criteria: *correctness* on the one side and *completeness* on the other side. An ideal learner is correct in the sense of making few mistakes, but also complete in the sense of abstaining but rarely. The two criteria are normally conflicting: increasing completeness typically comes along with reducing correctness and vice versa.

### 4.1 Correctness

As a measure of correctness, we propose a quantity that is also known as the *gamma rank correlation* [18] in statistics, although it is not applied to partial orders. Instead, it is used as a measure of correlation between rankings (with ties). As will be seen, however, it can also be used in a more general way.

Let $\sqsupset_*$ be the ground-truth relation on the set of alternatives $\mathbf{A}$. If this relation is a total order, like in label ranking, then $\boldsymbol{a} \sqsupset_* \boldsymbol{b}$ if $\boldsymbol{a}$ precedes $\boldsymbol{b}$ and $\boldsymbol{b} \sqsupset_* \boldsymbol{a}$ if $\boldsymbol{b}$ precedes $\boldsymbol{a}$; exactly one of these two cases is true, i.e., we never have $\boldsymbol{a} \perp_* \boldsymbol{b}$. Interestingly, in the case of instance ranking, it is not entirely clear whether the ground-truth is a total or a partial order. The goal of most learning algorithms for AUC maximization is to sort instances $\boldsymbol{x}$ according to their probability of belonging to the positive class, $\mathbf{P}(y = 1 \,|\, \boldsymbol{x})$.[4] Seen from this point of view, the underlying ground-truth is assumed to be a complete order. On the other hand, this complete order is never known and, therefore, can never be used as a reference for evaluating a prediction. Instead, only the class information is provided, and given a concrete test sample, evaluation measures like AUC do not care about the relative order of instances from the same class. In that sense, the ground-truth is treated like a partial order: $\boldsymbol{a} \sqsupset_* \boldsymbol{b}$ whenever $\boldsymbol{a}$ is positive and $\boldsymbol{b}$ negative (or, in the multi-class case, if the class of $\boldsymbol{a}$ is higher than the class of $\boldsymbol{b}$), while $\boldsymbol{a} \perp_* \boldsymbol{b}$ when $\boldsymbol{a}$ and $\boldsymbol{b}$ belong to the same class.

Now, let $\sqsupset$ be a predicted (strict) partial order, i.e., a prediction of $\sqsupset_*$. We call a pair of alternatives $\boldsymbol{a}$ and $\boldsymbol{b}$ *concordant* if they ought to be compared, because $\neg(\boldsymbol{a} \perp_* \boldsymbol{b})$, and are indeed compared in the correct way, that is,

$$(\boldsymbol{a} \sqsupset_* \boldsymbol{b} \wedge \boldsymbol{a} \sqsupset \boldsymbol{b}) \vee (\boldsymbol{b} \sqsupset_* \boldsymbol{a} \wedge \boldsymbol{b} \sqsupset \boldsymbol{a}) \ .$$

Likewise, we call $\boldsymbol{a}$ and $\boldsymbol{b}$ *discordant* if they ought to be compared, but the comparison is incorrect, that is,

$$(\boldsymbol{a} \sqsupset_* \boldsymbol{b} \wedge \boldsymbol{b} \sqsupset \boldsymbol{a}) \vee (\boldsymbol{b} \sqsupset_* \boldsymbol{a} \wedge \boldsymbol{a} \sqsupset \boldsymbol{b}) \ .$$

Note that, if $\boldsymbol{a} \perp_* \boldsymbol{b}$ (there is no need to compare $\boldsymbol{a}$ and $\boldsymbol{b}$) or $\boldsymbol{a} \perp \boldsymbol{b}$ (abstention on $\boldsymbol{a}$ and $\boldsymbol{b}$), then the two alternatives are neither concordant nor discordant.

[4] Indeed, this prediction maximizes the *expected* AUC on a test set.

Given these notions of concordance and discordance, we can define

$$\mathrm{CR}(\sqsupset, \sqsupset_*) \;=\; \frac{|C| - |D|}{|C| + |D|} \;, \tag{8}$$

where $C$ and $D$ denote, respectively, the set of concordant and discordant pairs of alternatives. Obviously, $\mathrm{CR}(\sqsupset, \sqsupset_*) = 1$ for $\sqsupset_*$=$\sqsupset$ and $\mathrm{CR}(\sqsupset, \sqsupset_*) = -1$ if $\sqsupset$ is the inversion of $\sqsupset_*$.

It is also interesting to mention that (8) is indeed a proper generalization of commonly used measures for the complete (non-partial) case, in the sense of reducing to these measures if $\sqsupset$ is a total order. In particular, it is easy to see that (8) reduces to Kendall's tau (1) in the case of label ranking (where $\sqsupset_*$ is a total order, too), and to the AUC measure (2) in the case of instance ranking (where $\sqsupset_*$ is a partial order).

### 4.2 Completeness

To measure the degree of completeness of a prediction, a straightforward idea is to punish the abstention from comparisons that should actually be made (while ignoring or, say, tolerating comparisons that are made despite not being necessary). This leads to the following measure of completeness:

$$\mathrm{CP}(\sqsupset) \;=\; \frac{|C| + |D|}{|\sqsupset_*|} \tag{9}$$

## 5 Experimental Results

As mentioned before, our method described in Section 3 can be applied to different ranking problems in a generic way. In this section, we present experimental results for two of the ranking problems outlined in Section 2, namely instance ranking and label ranking.

### 5.1 Instance Ranking

To test our method in the instance ranking scenario, we have selected a set of 16 binary classification data sets from the UCI repository and the Statlog collection[5]. We have used logistic regression as a base learner and produced ensembles of size 10.

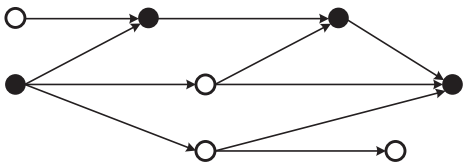
The values reported in Table 1 are averages over five repetitions of a 10-fold cross-validation. Comparing the correctness of predictions in terms of (8), it can be seen that our approach of partial abstention generally leads to improved performance. In fact, it is never worse and yields better results most of the time, sometimes with significant margins. Moreover, this gain in performance comes with an acceptable loss in terms of completeness. Indeed, the degrees of completeness are quite high throughout, often significantly above 90%.

**Table 1.** Results for instance ranking: mean values and standard deviations for correctness and completeness.

| data set | #attr. | #inst. | correctness with abstention | correctness w/o abstention | completeness |
|---|---|---|---|---|---|
| breast | 9 | 286 | 0.330±0.150 | 0.318±0.141 | 0.578±0.074 |
| breast-w | 9 | 699 | 0.988±0.014 | 0.987±0.015 | 0.982±0.015 |
| horse colic | 22 | 368 | 0.734±0.135 | 0.697±0.142 | 0.790±0.044 |
| credit rating | 15 | 690 | 0.858±0.062 | 0.827±0.065 | 0.888±0.038 |
| credit german | 20 | 1000 | 0.610±0.088 | 0.568±0.084 | 0.741±0.060 |
| pima diabetes | 8 | 768 | 0.684±0.084 | 0.666±0.086 | 0.819±0.047 |
| heart statlog | 13 | 270 | 0.811±0.102 | 0.797±0.101 | 0.890±0.060 |
| hepatitis | 19 | 155 | 0.709±0.292 | 0.697±0.271 | 0.797±0.084 |
| ionosphere | 34 | 351 | 0.771±0.174 | 0.722±0.190 | 0.814±0.098 |
| kr-vs-kp | 36 | 3196 | 0.992±0.006 | 0.980±0.007 | 0.991±0.006 |
| labor | 16 | 57 | 0.990±0.049 | 0.985±0.060 | 0.989±0.052 |
| mushroom | 22 | 8124 | 1.000±0.000 | 1.000±0.000 | 0.808±0.017 |
| thyroid disease | 29 | 3772 | 0.890±0.071 | 0.883±0.070 | 0.928±0.040 |
| sonar | 60 | 206 | 0.684±0.224 | 0.575±0.271 | 0.575±0.056 |
| tic-tac-toe | 9 | 958 | 0.253±0.127 | 0.221±0.120 | 0.908±0.013 |
| vote | 16 | 435 | 0.981±0.032 | 0.976±0.036 | 0.913±0.035 |

We conducted a second experiment with the aim to investigate the trade-off between correctness and completeness. As was mentioned earlier, and to some extent already confirmed by our first experiment, we expect a compromise between both criteria insofar as it should be possible to increase correctness at the cost of completeness. To verify this conjecture, we varied the threshold $\alpha$ in (4) in the range $[\alpha^*, 1]$. Compared to the use of $\alpha^*$, larger thresholds will make the predictions increasingly incomplete; at the same time, however, they should also become more correct. Indeed, the results we obtained are well in agreement with these expectations. Fig. 2 shows typical examples of the trade-off between correctness and completeness for two data sets.

Finally, it is interesting to look at the maximal chains of a predicted partial order. A maximal chain of a partially ordered set $X$ is a maximal subset $C \subset X$ that is totally ordered, like the set of elements depicted as black nodes in the following partial order:



The remaining elements $X \setminus C$ can then be considered as those that cannot be inserted into the order in a reliable way, and are hence ignored. Since each maximal chain is a total order, it can be visualized in terms of an ROC curve.

[5] `www.ics.uci.edu/~mlearn/MLRepository.html`, `lib.stat.cmu.edu/`

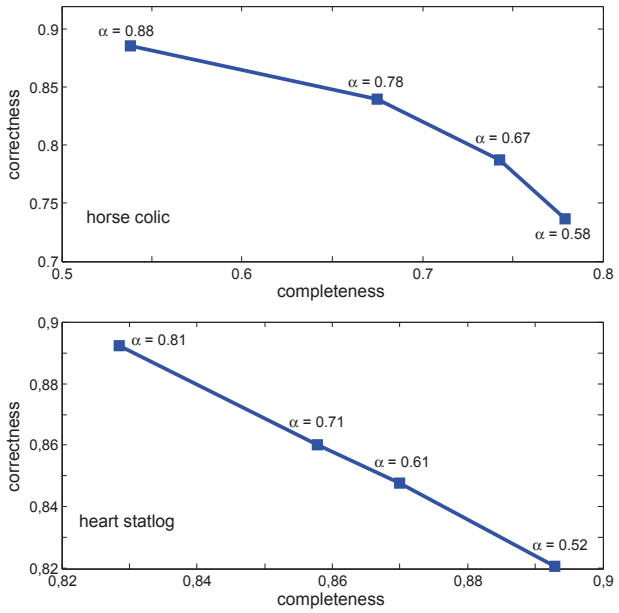



**Fig. 2.** Instance ranking with partial abstention: Trade-off between correctness and completeness for selected data sets.

A typical example for the sonar data set is shown in Fig. 3. As can be seen, the curves for the maximal chains tend to dominate the original ROC curve for this data, suggesting that the ranking of elements in the individual chains is indeed more reliable than the ranking of the complete data.

### 5.2 Label Ranking

In view of a lack of benchmark data for label ranking, we resorted to multi-class data sets from the UCI repository and turned them into label ranking data by following the procedure proposed in [15]: A naive Bayes classifier is first trained on the complete data set. Then, for each example, all the labels present in the data set are ordered with respect to the predicted class probabilities (in the case of ties, labels with lower index are ranked first).[6]

The setting of this experiment is similar to the one we did for instance ranking. We performed five repetitions of a 10-fold cross-validation and used an ensemble size of 10. As a label ranking method, we used the *ranking by pairwise comparison* approach [15], again with logistic regression as a base learner.

The results, summarized in Table 2, convey the same message as before: Correctness can be improved at the cost of completeness, and compared to the case of instance ranking, the loss in completeness is even much smaller here; see

[6] The data sets are available at `www.uni-marburg.de/fb12/kebi/research`.

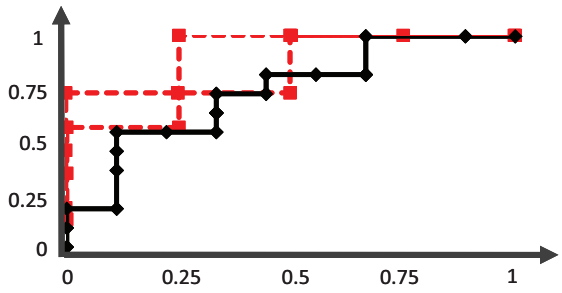



**Fig. 3.** ROC curves for the maximal chains of a predicted partial order (dashed lines) and the complete data (solid line) for the sonar data.

also Fig. 4, in which we show the same kind of trade-off curves as for the case of instance ranking.

**Table 2.** Results for label ranking: mean values and standard deviations for correctness and completeness.

| data set | #attr. | #classes | #inst. | correctness with abstention | correctness w/o abstention | completeness |
|---|---|---|---|---|---|---|
| iris | 4 | 3 | 150 | 0.910±0.062 | 0.885±0.068 | 0.991±0.063 |
| wine | 13 | 3 | 178 | 0.940±0.051 | 0.921±0.053 | 0.988±0.067 |
| glass | 9 | 6 | 214 | 0.892±0.039 | 0.882±0.042 | 0.990±0.030 |
| vowel | 10 | 11 | 528 | 0.657±0.019 | 0.647±0.019 | 0.988±0.016 |
| vehicle | 18 | 4 | 846 | 0.858±0.026 | 0.854±0.025 | 0.992±0.039 |
| authorship | 70 | 4 | 841 | 0.941±0.016 | 0.910±0.015 | 0.989±0.043 |
| pendigits | 16 | 10 | 10992 | 0.933±0.002 | 0.932±0.002 | 0.999±0.005 |
| segment | 18 | 7 | 2310 | 0.938±0.006 | 0.934±0.006 | 0.998±0.011 |

## 6 Conclusions and Future Work

In this paper, we have addressed the problem of "reliable" prediction in the context of learning to rank. In this regard, we have made the following main contributions:

- Based on the idea of allowing a learner to abstain from an uncertain comparison of alternatives, together with the requirement that predictions are consistent, we have proposed a relaxation of the conventional setting in which predictions are given in terms of partial instead of total orders.
- We have proposed a generic approach to predicting partial orders or, according to our interpretation, ranking with partial abstention, which is applicable to different types of ranking problems.

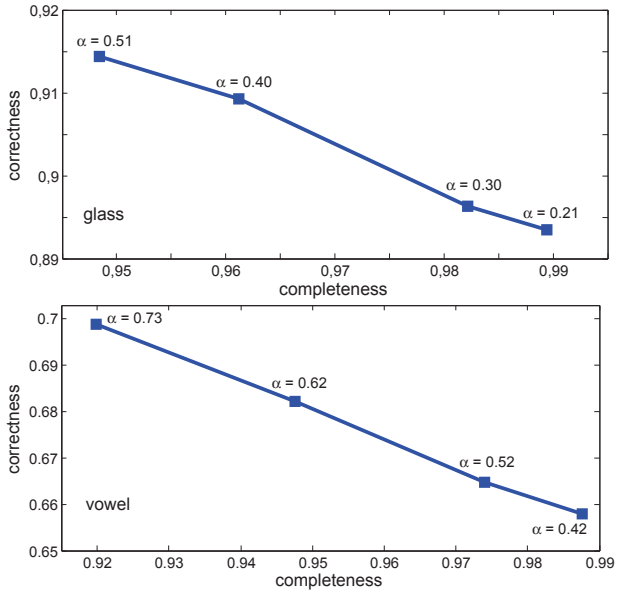



**Fig. 4.** Label ranking with partial abstention: Trade-off between correctness and completeness for selected data sets.

- We have introduced reasonable measures for evaluating the performance of a ranker with (partial) reject option, namely measures of correctness and completeness. These measures are proper generalizations of conventional and commonly used measures for total orders.
- Empirically, we have shown that our method is indeed able to trade off accuracy against completeness: The correctness of a prediction can be increased at the cost of reducing the number of alternatives that are compared.

The extension from predicting total to predicting partial orders as proposed in this paper opens the door for a multitude of further studies. Here, we mention just one example of an interesting direction for future work, which concerns the type of target order $\sqsupset_*$ to be predicted. In this paper, we have essentially assumed that the target is a complete order, and a prediction in terms of a partial order $\sqsupset$ an incomplete estimation thereof, even though it was already mentioned that, in the case of instance ranking (AUC maximization), the target may also be considered as a partial order. However, even in that case, our evaluation measure does not penalize the prediction of an order relation between two instances $\boldsymbol{a}$ and $\boldsymbol{b}$ from the same class. In other words, we do not penalize the case where $\boldsymbol{a} \sqsupset \boldsymbol{b}$ even though $\boldsymbol{a} \perp_* \boldsymbol{b}$. Now, if $\sqsupset_*$ is a true partial order, it clearly makes sense to request, not only the correct prediction of order relations $\boldsymbol{a} \sqsupset_* \boldsymbol{b}$ between alternatives, but also of incomparability relations $\boldsymbol{a} \perp_* \boldsymbol{b}$. Although the difference may look subtle at first sight, the changes will go beyond the evaluation of predictions and instead call for different learning algorithms. In particular, in

this latter scenario, $\boldsymbol{a} \perp \boldsymbol{b}$ will be interpreted as a prediction that $\boldsymbol{a}$ and $\boldsymbol{b}$ are incomparable ($\boldsymbol{a} \perp_* \boldsymbol{b}$), and not as a rejection of the decision whether $\boldsymbol{a} \sqsupset_* \boldsymbol{b}$ or $\boldsymbol{b} \sqsupset_* \boldsymbol{a}$. Nevertheless, the two settings are of course related, and we plan to elaborate on their connection in future work.

## References


1. Bakir, G., Hofmann, T., Schölkopf, B., Smola, A., Taskar, B., Vishwanathan, S., eds.: Predicting structured data. MIT Press (2007)
2. Chow, C.: On optimum recognition error and reject tradeoff. IEEE Transactions on Information Theory **16**(1) (1970) 41–46
3. Herbei, R., Wegkamp, M.H.: Classification with reject option. Canadian Journal of Statistics **34**(4) (2006) 709–721
4. Bartlett, P.L., Wegkamp, M.H.: Classification with a reject option using a hinge loss. Journal of Machine Learning Research **9** (2008) 1823–1840
5. Fürnkranz, J., Hüllermeier, E., eds.: Preference Learning. Springer-Verlag (2010)
6. Har-Peled, S., Roth, D., Zimak, D.: Constraint classification for multiclass classification and ranking. In Becker, S., Thrun, S., Obermayer, K., eds.: Advances in Neural Information Processing Systems 15 (NIPS-02). (2003) 785–792
7. Fürnkranz, J., Hüllermeier, E.: Pairwise preference learning and ranking. In: Proc. ECML–03, 13th European Conference on Machine Learning, Cavtat-Dubrovnik, Croatia (September 2003)
8. Dekel, O., Manning, C., Singer, Y.: Log-linear models for label ranking. In: Advances in Neural Information Processing Systems. (2003)
9. Fürnkranz, J., Hüllermeier, E., Vanderlooy, S.: Binary decomposition methods for multipartite ranking. In: Proceedings ECML/PKDD–2009, European Conference on Machine Learning and Knowledge Discovery in Databases, Bled, Slovenia (2009)
10. Cohen, W., Schapire, R., Singer, Y.: Learning to order things. Journal of Artificial Intelligence Research **10** (1999) 243–270
11. Fawcett, T.: An introduction to ROC analysis. Pattern Recognition Letters **27**(8) (2006) 861–874
12. Gnen, M., Heller, G.: Concordance probability and discriminatory power in proportional hazards regression. Biometrika **92**(4) (2005) 965–970
13. Fagin, R., Kumar, R., Sivakumar, D.: Comparing top k lists. SIAM Journal of Discrete Mathematics **17**(1) (2003) 134–160
14. Manning, C., Raghavan, P., Schütze, H.: Introduction to Information Retrieval. Cambridge University Press (2008)
15. Hüllermeier, E., Fürnkranz, J., Cheng, W., Brinker, K.: Label ranking by learning pairwise preferences. Artificial Intelligence **172** (2008) 1897–1917
16. Rademaker, M., De Baets, B.: A threshold for majority in the context of aggregating partial order relations. In: Proc. WCCI–2010, World Congress on Computational Intelligence, Barcelona, Spain (2010)
17. Floyd, R.: Algorithm 97: Shortest path. Communications of the ACM **5** (1962)
18. Goodman, L., Kruskal, W.: Measures of Association for Cross Classifications. Springer-Verlag, New York (1979)