Ranking Forests

Stéphan Clémençon STEPHAN.CLEMENCON@TELECOM-PARISTECH.FR
Marine Depecker MARINE-DEPECKER@TELECOM-PARISTECH.FR
Institut Telecom LTCI - UMR Telecom ParisTech/CNRS No. 5141
Telecom ParisTech
46 rue Barrault, Paris, 75634, France

Nicolas Vayatis NICOLAS.VAYATIS@CMLA.ENS-CACHAN.FR
CMLA - UMR ENS Cachan/CNRS No. 8536
ENS Cachan
61, avenue du Président Wilson, Cachan, 94230, France

Editor: Tong Zhang

Abstract

The present paper examines how the aggregation and feature randomization principles underlying the algorithm RANDOM FOREST (Breiman, 2001) can be adapted to *bipartite ranking*. The approach taken here is based on nonparametric scoring and ROC curve optimization in the sense of the AUC criterion. In this problem, aggregation is used to increase the performance of scoring rules produced by ranking trees, as those developed in Clémençon and Vayatis (2009c). The present work describes the principles for building median scoring rules based on concepts from rank aggregation. Consistency results are derived for these aggregated scoring rules and an algorithm called RANKING FOREST is presented. Furthermore, various strategies for feature randomization are explored through a series of numerical experiments on artificial data sets.

Keywords: bipartite ranking, nonparametric scoring, classification data, ROC optimization, AUC criterion, tree-based ranking rules, bootstrap, bagging, rank aggregation, median ranking, feature randomization

1. Introduction

Aggregating decision rules or function estimators has now become a folk concept in machine learning and nonparametric statistics. Indeed, the idea of combining decision rules with an additional randomization ingredient brings a dramatic improvement of performance in various contexts. These ideas go back to the seminal work of Amit and Geman (1997), Breiman (1996), and Nemirovski (2000). However, in the context of the "learning-to-rank" problem, the implementation of this idea is still at a very early stage. In the present paper, we propose to take one step beyond in the program of boosting performance by aggregation and randomization for this problem. The present paper explores the particular case of learning to rank high dimensional observation vectors in presence of binary feedback. This case is also known as the bipartite ranking problem, see Freund et al. (2003), Agarwal et al. (2005), Clémençon et al. (2005). The setup of bipartite ranking is useful when considering real-life applications such as credit-risk or medical screening, spam filtering, or recommender systems. There are two major approaches to bipartite ranking: the *preference-based* approach (see Cohen et al. 1999) and the *scoring-based* approach (in the spirit of logistic regression methods, see, e.g., Hastie and Tibshirani 1990, Hilbe 2009). The idea of combining ranking rules to learn preferences was introduced in Freund et al. (2003) with a boosting algorithm and the consistency for this type of methods was proved in Clémençon et al. (2008) by reducing the bipartite ranking problem to a classification problem over pairs of observations (see also Agarwal et al. 2005). Here, we will cast bipartite ranking in the context of nonparametric scoring and we will consider the issue of combining randomized *scoring rules*. Scoring rules are real-valued functions mapping the observation space with the real line, thus conveying an order relation between high dimensional observation vectors.

Nonparametric scoring has received an increasing attention in the machine learning literature as a part of the growing interest which affects ROC analysis. The scoring problem can be seen as a learning problem where one observes input observation vectors X in a high dimensional space $\mathcal{X}$ and receives only a binary feedback information through an output variable $Y \in \{-1,+1\}$. Whereas classification only focuses on predicting the label $\widetilde{Y}$ of a new observation $\widetilde{X}$, scoring algorithms aim at recovering an order relation on $\mathcal{X}$ in order to predict the ordering over a new sample of observation vectors $X'_1,\ldots,X'_m$ so that there as many as possible positive instances at the top of the list. From a statistical perspective, the scoring problem is more difficult than classification but easier than regression. Indeed, in classification, the goal is to learn *one* single level set of the regression function whereas, in scoring, one wants to recover the nested collection of *all* the level sets of the regression function (without necessarily knowing the corresponding levels), but not the regression function itself (see Clémençon and Vayatis 2009b). In previous work, we developed a tree-based procedure for nonparametric scoring called TREERANK, see Clémençon and Vayatis (2009c), Clémençon et al. (2010). The TREERANK algorithm and its variants produce scoring rules expressed as partitions of the input space coupled with a permutation over the cells of the partition. These scoring rules present the interesting feature that they can be stored in an oriented binary tree structure, called a *ranking tree*. Moreover, their very construction actually implements the optimization of the ROC curve which reflects the quality measure of the scoring rule for the end-user.

The use of resampling in this context was first considered in Clémençon et al. (2009). A more thorough analysis is developed throughout this paper and we show how to combine feature randomization and bootstrap aggregation techniques based on the ranking trees produced by the TREERANK algorithm in order to increase ranking performance in the sense of the ROC curve. In the classification setup, theoretical evidence has been recently provided for the aggregation of randomized classifiers in the spirit of random forests (see Biau et al. 2008). However, in the context of ROC optimization, combining scoring rules through naive aggregation does not necessarily make sense. Our approach builds on the advances in the rank aggregation problem. Rank aggregation was originally introduced in social choice theory (see Barthélémy and Montjardet 1981 and the references therein) and recently "rediscovered" in the context of internet applications (see Pennock et al. 2000). For our needs, we shall focus on *metric-based consensus methods* (see Hudry 2004 or Fagin et al. 2006, and the references therein), which provide the key to the aggregation of ranking trees. In the paper, we also discuss various aspects of feature randomization which can be incorporated at various levels in ranking trees. Also a novel ranking methodology, called RANKING FOREST, is introduced.

The article is structured as follows. Section 2 sets out the notations and shortly describes the main notions for the bipartite ranking problem. Section 3 describes the elements from the theory of rank aggregation and measures of consensus leading to the aggregation of scoring rules defined over finite partitions of the input space. The next section presents the main theoretical results of the paper which are consistency results for scoring rules based on the aggregation of randomized piecewise constant scoring rules. Section 5 presents RANKING FOREST, a new algorithm for nonparametric scoring which implements the theoretical concepts developed so far. Section 6 presents an empirical study of the RANKING FOREST algorithm with numerical results based on simulated data. Finally, some concluding remarks are collected in Section 7. Reminders, technical details and proofs are deferred to the Appendix.

2. Probabilistic Setup for Bipartite Ranking

ROC analysis is a popular way of evaluating the capacity of a given scoring rule to discriminate between two populations, see Egan (1975). ROC curves and related performance measures such as the AUC have now become of standard use for assessing the quality of ranking methods in a bipartite framework. Throughout this section, we recall basic concepts related to bipartite ranking from the angle of ROC analysis.

Modeling the data. The probabilistic setup is the same as in standard binary classification. The random variable Y is a binary label, valued in $\{-1,+1\}$, while the random vector $X=(X^{(1)},\ldots,X^{(q)})$ models some multivariate observation for predicting Y, taking its values in a high-dimensional space $\mathcal{X}\subset\mathbb{R}^q$, $q\geq 1$. The probability measure on the underlying space is entirely described by the pair (μ,η), where μ denotes the marginal distribution of X and $\eta(x)=\mathbb{P}\{Y=+1\mid X=x\}$, $x\in\mathcal{X}$, the posterior probability. With no restriction, here we assume that $\mathcal{X}$ coincides with the support of μ.

The scoring approach to bipartite ranking. An informal way of considering the ranking task under this model is as follows. Given a a sample of independent copies of the pair (X,Y), the goal is to learn how to order new data $X_1,\ldots,X_m$ without label feedback, so that positive instances are mostly at the top of the resulting list with large probability. A natural way of defining a total order on the multidimensional space $\mathcal{X}$ is to map it with the natural order on the real line by means of a *scoring rule*, that is, a measurable mapping $s:\mathcal{X}\to\mathbb{R}$. A preorder[1] $\preccurlyeq_s$ on $\mathcal{X}$ is then defined by: $\forall(x,x')\in\mathcal{X}^2$, $x\preccurlyeq_s x'$ if and only if $s(x)\leq s(x')$.

Measuring performance. The capacity of a candidate s to discriminate between the positive and negative populations is generally evaluated by means of its ROC curve (standing for "Receiver Operating Characteristic" curve), a widely used functional performance measure which we recall here.

Definition 1 (TRUE ROC CURVE) *Let s be a scoring rule. The true* ROC *curve of s is the "probability-probability" plot given by:*

$$t\in\mathbb{R}\mapsto\left(\mathbb{P}\left\{s(X)>t\mid Y=-1\right\},\mathbb{P}\left\{s(X)>t\mid Y=1\right\}\right)\in[0,1]^2\ .$$

By convention, when a jump occurs in the plot of the ROC *curve, the corresponding extremities of the curve are connected by a line segment, so that the* ROC *curve of s can be viewed as the graph of a continuous mapping* $\alpha\in[0,1]\mapsto\mathrm{ROC}(s,\alpha)$.

We refer to Clémençon and Vayatis (2009c) for a list of properties of ROC curves (see the Appendix section therein). The ROC curve offers a visual tool for assessing ranking performance (see Figure 1): the closer to the left upper corner of the unit square $[0,1]^2$ the curve $\mathrm{ROC}(s,.)$, the better the scoring rule s. Therefore, the ROC curve conveys a partial order on the set of all

[1] A preorder is a binary relation which is reflexive and transitive.

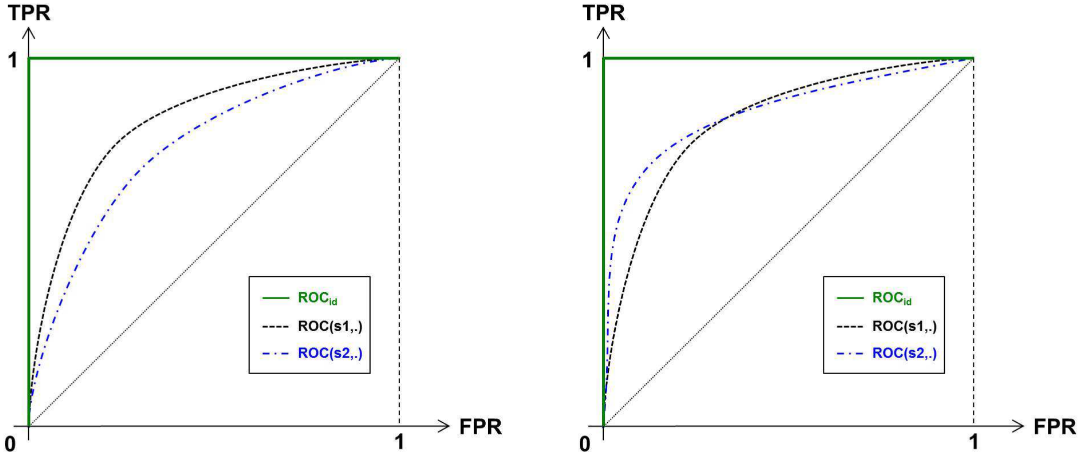


Figure 1: ROC curves.

scoring rules: for all pairs of scoring rules s_1 and s_2, we say that s_2 is more accurate than s_1 when $\mathrm{ROC}(s_1,\alpha) \leq \mathrm{ROC}(s_2,\alpha)$ for all $\alpha \in [0,1]$. By a standard Neyman-Pearson argument, one may establish that the most accurate scoring rules are increasing transforms of the regression function which is equal to the conditional probability function η up to an affine transformation.

Definition 2 (OPTIMAL SCORING RULES) *We call optimal scoring rules the elements of the set $\mathcal{S}^*$ of scoring functions s^* such that $\forall (x,x') \in \mathcal{X}^2$, $\eta(x) < \eta(x') \Rightarrow s^*(x) < s^*(x')$.*

The fact that the elements of $\mathcal{S}^*$ are optimizers of the ROC curve is shown in Clémençon and Vayatis (2009c) (see Proposition 4 therein). When, in addition, the random variable $\eta(X)$ is assumed to be continuous, then $\mathcal{S}^*$ coincides with the set of strictly increasing transforms of η. The performance of a candidate scoring rule s is often summarized by a scalar quantity called the *Area Under the* ROC *Curve* (AUC) which can be considered as a summary of the ROC curve. In the paper, we shall use the following definition of the AUC.

Definition 3 (AUC) *Let s be a scoring rule. The* AUC *is the functional defined as:*

$$\mathrm{AUC(s)} = \mathbb{P}\{\mathrm{s(X_1) < s(X_2) \mid (Y_1,Y_2) = (-1,+1)}\}$$
$$+\frac{1}{2}\mathbb{P}\{s(X_1) = s(X_2) \mid (Y_1,Y_2) = (-1,+1)\},$$

where (X_1,Y_1) and (X_2,Y_2) denote two independent copies of the pair (X,Y), for any scoring function s.

This functional provides a *total order* on the set of scoring rules and, equipped with the convention introduced in Definition 1, AUC(s) coincides with $\int_0^1 \mathrm{ROC}(\mathrm{s},\alpha)\,\mathrm{d}\alpha$ (see, for instance, Proposition 1 in Clémençon et al. 2011). We shall denote the optimal curve and the corresponding (maximum) value for the AUC criterion by $\mathrm{ROC}^* = \mathrm{ROC}(\mathrm{s}^*,.)$ and $\mathrm{AUC}^* = \mathrm{AUC}(\mathrm{s}^*)$, where $s^* \in \mathcal{S}^*$. The

statistical counterparts of ROC(s, .) and AUC(s) based on sampling data $\mathcal{D}_n = \{(X_i, Y_i) : 1 \leq i \leq n\}$ are obtained by replacing the class distributions by their empirical versions in the definitions. They are denoted by $\widehat{\mathrm{ROC}}(s, .)$ and $\widehat{\mathrm{AUC}}(s)$ in the sequel.
Piecewise constant scoring rules. In the paper, we will focus on a particular subclass of scoring rules.

Definition 4 (PIECEWISE CONSTANT SCORING RULE) *A scoring rule s is piecewise constant if there exists a finite partition $\mathcal{P}$ of $\mathcal{X}$ such that for all $\mathcal{C} \in \mathcal{P}$, there exists a constant $k_{\mathcal{C}} \in \mathbb{R}$ such that $\forall x \in \mathcal{C}$, $s(x) = k_{\mathcal{C}}$.*

This definition does not provide a unique characterization of the underlying partition. The partition $\mathcal{P}$ is minimal if, for any two of its elements $\mathcal{C} \neq \mathcal{C}'$, we have $k_{\mathcal{C}} \neq k_{\mathcal{C}'}$. The scoring rule conveys an ordering on the cells of the minimal partition.

Definition 5 (RANK OF A CELL) *Let s be a scoring rule and $\mathcal{P}$ the associated minimal partition. The scoring rule induces a ranking $\preceq_s$ over the cells of the partition. For a given cell $\mathcal{C} \in \mathcal{P}$, we define its rank $\mathcal{R}_{\preceq_s}(\mathcal{C}) \in \{1, \ldots, |\mathcal{P}|\}$ as the rank affected by the ranking $\preceq_s$ over the elements of the partition. By convention, we set rank 1 to correspond to the highest score.*

The advantage of the class of piecewise constant scoring rules is that they provide finite rankings on the elements of $\mathcal{X}$ and they will be the key for applying the aggregation procedure.

3. Aggregation of Scoring Rules

In recent years, the issue of summarizing or aggregating various rankings has been a topic of growing interest in the machine-learning community. This evolution was mainly motivated by practical problems in the context of internet applications: design of meta-search engines, collaborative filtering, spam-fighting, *etc.* We refer for instance to Pennock et al. (2000), Dwork et al. (2001), Fagin et al. (2003) and Ilyas et al. (2002). Such problems have led to a variety of results, ranging from the generalization of the mathematical concepts introduced in social choice theory (see Barthélémy and Montjardet 1981 and the references therein) for defining relevant notions of *consensus* between rankings (Fagin et al., 2006), to the development of efficient procedures for computing such "consensus rankings" (Betzler et al., 2008; Mandhani and Meila, 2009; Meila et al., 2007) through the study of probabilistic models over sets of rankings (Fligner and Verducci , Eds.; Lebanon and Lafferty, 2003). Here we consider rank aggregation methods in the perspective of extending the bagging approach to ranking trees.

3.1 The Case of Piecewise Constant Scoring Rules

The ranking rules considered in this paper result from the aggregation of a collection of piecewise constant scoring rules. Since each of these scoring rules is related to a possibly different partition, we are lead to consider a collection of partitions of $\mathcal{X}$. Hence, the aggregated rule needs to be defined on the least fine subpartition of this collection of partitions.

Definition 6 (SUBPARTITION OF A COLLECTION OF PARTITIONS) *Consider a collection of B partitions of $\mathcal{X}$ denoted by $\mathcal{P}_b$, $b = 1, \ldots, B$. A subpartition of this collection is a partition $\mathcal{P}_B$ made* *of nonempty subsets $\mathcal{C} \subset \mathcal{X}$ which satisfy the following constraint : for all $\mathcal{C} \in \mathcal{P}_B$, there exists $(\mathcal{C}_1, \ldots, \mathcal{C}_B) \in \mathcal{P}_1 \times \cdots \times \mathcal{P}_B$ such that*

$$\mathcal{C} \subseteq \bigcap_{b=1}^{B} \mathcal{C}_b \, .$$

We denote $\mathcal{P}_B^* = \bigcap_{b \leq B} \mathcal{P}_b$.

One may easily see that $\mathcal{P}_B^*$ is a subpartition of any of the $\mathcal{P}_b$'s, and the largest one in the sense that any partition $\mathcal{P}$ which is a subpartition of $\mathcal{P}_b$ for all $b \in \{1, \ldots, B\}$ is a subpartition of $\mathcal{P}_B^*$. The case where the partitions are obtained from a binary tree structure is of particular interest as we shall consider tree-based piecewise constant scoring rules later on. Incidentally, it should be noticed that, from a computational perspective, the underlying tree structures considerably help in getting the cells of $\mathcal{P}_B^*$ explicitly. We refer to Appendix D for further details.

Now consider a collection of piecewise constant scoring rules s_b, $b = 1, \ldots, B$, and denote their associated (minimal) partitions by $\mathcal{P}_b$. Each scoring rule s_b naturally induces a ranking (or a *preorder*) $\preceq_b^*$ on the partition $\mathcal{P}_B^*$. Indeed, for all $(\mathcal{C}, \mathcal{C}') \in \mathcal{P}_B^{*2}$, one writes by definition $\mathcal{C} \preceq_b^* \mathcal{C}'$ (respectively, $\mathcal{C} \prec_b^* \mathcal{C}'$) if and only if $\mathcal{C}_b \preceq_b^* \mathcal{C}_b'$ (respectively, $\mathcal{C}_b \prec_b^* \mathcal{C}_b'$) where $(\mathcal{C}_b, \mathcal{C}_b') \in \mathcal{P}_b^2$ are such that $\mathcal{C} \times \mathcal{C}' \subset \mathcal{C}_b \times \mathcal{C}_b'$.

The collection of scoring rules leads to a collection of B rankings on $\mathcal{P}_B^*$. Such a collection is called a *profile* in voting theory. Now, based on this *profile*, we would like to define a "central ranking" or a *consensus*. Whereas the mean, or the median, naturally provides such a summary when considering scalar data, various meanings can be given to this notion for rankings (see Appendix B).

3.2 Probabilistic Measures of Scoring Agreement

The purpose of this subsection is to extend the concept of measures of agreement for rankings to scoring rules defined over a general space $\mathcal{X}$ which is not necessarily finite. In practice, however, we will only consider the case of piecewise constant scoring rules and we shall rely on the definition of the probabilistic Kendall tau.

Notations. We already introduced the notation $\preceq_s$ for the preorder relation over the cells of a partition $\mathcal{P}$ as induced by a piecewise scoring rule s. We shall use the 'curly' notation for the preorder relation $\preccurlyeq_s$ on $\mathcal{X}$ which is described through the following condition: $\forall \mathcal{C}, \mathcal{C}' \in \mathcal{P}$, we have $x \preccurlyeq_s x'$, $\forall x \in \mathcal{C}$, $\forall x' \in \mathcal{C}'$, if and only if $\mathcal{C} \preceq_s \mathcal{C}'$. This is also equivalent to $s(x) \leq s(x')$, $\forall x \in \mathcal{C}$, $\forall x' \in \mathcal{C}'$. We now introduce a measure of similarity for preorders on $\mathcal{X}$ induced by scoring rules s_1 and s_2.

We recall here the definition of the theoretical Kendall τ between two random variables.

Definition 7 (PROBABILISTIC KENDALL τ) *Let (Z_1, Z_2) be two random variables defined on the same probability space. The probabilistic Kendall τ is defined as*

$$\tau(Z_1, Z_2) = 1 - 2 d_\tau(Z_1, Z_2) \, ,$$

with:

$$d_\tau(Z_1, Z_2) = \mathbb{P}\{(Z_1 - Z_1') \cdot (Z_2 - Z_2') < 0\} + \frac{1}{2}\mathbb{P}\{Z_1 = Z_1', \, Z_2 \neq Z_2'\} + \frac{1}{2}\mathbb{P}\{Z_1 \neq Z_1', \, Z_2 = Z_2'\}.$$

where (Z_1', Z_2') *is an independent copy of the pair* (Z_1, Z_2).

As shown by the following result, whose proof is left to the reader, the Kendall τ for the pair $(s(X), Y)$ is related to AUC(s).

Proposition 8 *We use the notation* $p = \mathbb{P}\{Y = 1\}$. *For any real-valued scoring rule s, we have:*

$$\frac{1}{2}\,(1-\tau(s(X),Y)) = 2p(1-p)\,(1-\mathrm{AUC(s)}) + \frac{1}{2}\mathbb{P}\{s(X) \neq s(X')\ ,\ Y = Y'\}\ .$$

For given scoring rules s_1 and s_2 and considering the probabilistic Kendall tau for random variables $s_1(X)$ and $s_2(X)$, we can set: $d_X(s_1, s_2) = d_\tau(s_1(X), s_2(X))$. One may easily check that d_X defines a distance between the orderings $\preccurlyeq_{s_1}$ and $\preccurlyeq_{s_2}$ induced by s_1 and s_2 on the set $\mathcal{X}$ (which is supposed to coincide with the support of the distribution of X). The following proposition shows that the deviation between scoring rules in terms of AUC is controlled by a quantity involving the probabilistic agreement based on Kendall tau.

Proposition 9 *(*AUC AND KENDALL τ*) Assume* $p \in (0,1)$. *For any scoring rules* s_1 *and* s_2 *on* $\mathcal{X}$, *we have:*

$$|\mathrm{AUC}(s_1) - \mathrm{AUC}(s_2)| \leq \frac{d_X(s_1, s_2)}{2p(1-p)} = \frac{1-\tau_X(s_1, s_2)}{4p(1-p)}\ .$$

The converse inequality does not hold in general. Indeed, scoring rules with same AUC may yield to different rankings. However, the following result guarantees that a scoring rule with a nearly optimal AUC is close to the optimal scoring rules in the sense of Kendall tau, under the additional assumption that the noise condition introduced in Clémençon et al. (2008) is fulfilled.

Proposition 10 *(*KENDALL τ AND OPTIMAL AUC*) Assume that the random variable* $\eta(X)$ *is continuous and that there exist* $c < \infty$ *and* $a \in (0,1)$ *such that:*

$$\forall x \in \mathcal{X},\ \ \mathbb{E}\left[|\eta(X) - \eta(x)|^{-a}\right] \leq c\ . \tag{1}$$

Then, we have, for any scoring rule s and any optimal scoring rule $s^* \in \mathcal{S}^*$:

$$1 - \tau_X(s^*, s) \leq C \cdot (\mathrm{AUC}^* - \mathrm{AUC(s)})^{a/(1+a)}\ ,$$

with $C = 3 \cdot c^{1/(1+a)} \cdot (2p(1-p))^{a/(1+a)}$.

Remark 11 *(*ON THE NOISE CONDITION*) As shown in previous work, the condition (1) is rather weak. Indeed, it is fulfilled for any* $a \in (0,1)$ *as soon the probability density function of* $\eta(X)$ *is bounded (see Corollary 8 in Clémençon et al. 2008).*

The next result shows the connection between the Kendall tau distance between preorders on $\mathcal{X}$ induced by piecewise constant scoring rules s_1 and s_2 and a specific notion of distance between the rankings $\preceq_{s_1}$ and $\preceq_{s_2}$ on $\mathcal{P}$.

Lemma 12 *Let* s_1, s_2, *two piecewise constant scoring rules. We have:*

$$d_X(s_1, s_2) = 2 \sum_{1 \leq k < l \leq K} \mu(\mathcal{C}_k)\mu(\mathcal{C}_l) \cdot U_{k,l}(\preceq_{s_1}, \preceq_{s_2})\ , \tag{2}$$

where, for two orderings $\preceq$, $\preceq'$ on a partition of cells $\{C_k \ : \ k=1,\dots,K\}$, we have:

$$\begin{aligned}U_{k,l}(\preceq,\preceq') = \mathbb{I}\{(\mathcal{R}_{\preceq}(\mathcal{C}_k)-\mathcal{R}_{\preceq}(\mathcal{C}_l))(\mathcal{R}_{\preceq'}(\mathcal{C}_k)-\mathcal{R}_{\preceq'}(\mathcal{C}_l))<0\}\\ +\frac{1}{2}\mathbb{I}\{\mathcal{R}_{\preceq}(\mathcal{C}_k)=s_{\preceq}(\mathcal{C}_l),\ \mathcal{R}_{\preceq'}(\mathcal{C}_k)\neq\mathcal{R}_{\preceq'}(\mathcal{C}_l)\}\\ +\frac{1}{2}\mathbb{I}\{\mathcal{R}_{\preceq'}(\mathcal{C}_k)=\mathcal{R}_{\preceq'}(\mathcal{C}_l),\ \mathcal{R}_{\preceq}(\mathcal{C}_k)\neq\mathcal{R}_{\preceq}(\mathcal{C}_l)\}\ .\end{aligned}$$

The proof is straightforward and thus omitted.

Notice that the term $U_{k,l}(\preceq_{s_1},\preceq_{s_2})$ involved in Equation (2) is equal to 1 when the cells $\mathcal{C}_k$ and $\mathcal{C}_l$ are not sorted in the same order by s_1 and s_2 (in absence of ties), to $1/2$ when they are tied for one ranking but not for the other, and to 0 otherwise. As a consequence, the agreement $\tau_X(s_1,s_2)$ may be viewed as a "weighted version" of the rate of concordant pairs of the cells of $\mathcal{P}$ measured by the classical Kendall τ (see the Appendix B). A statistical version of $\tau_X(s_1,s_2)$ is obtained by replacing the values of $\mu(\mathcal{C}_k)$ by their empirical counterparts in Equation (2). We thus set:

$$\widehat{\tau}_X(s_1,s_2)=1-2\widehat{d}_X(s_1,s_2), \tag{3}$$

where $\widehat{d}_X(s_1,s_2)=2/(n(n-1))\sum_{i<j}K(X_i,X_j)$ is a U-statistic of degree 2 with symmetric kernel given by:

$$\begin{aligned}K(x,x')=\mathbb{I}\{(s_1(x)-s_1(x'))\cdot(s_2(x)-s_2(x'))<0\}+\frac{1}{2}\mathbb{I}\{s_1(x)=s_1(x'),\ s_2(x)\neq s_2(x')\}\\ +\frac{1}{2}\mathbb{I}\{s_1(x)\neq s_1(x'),\ s_2(x)=s_2(x')\}\ .\end{aligned}$$

Remark 13 *Other measures of agreement between $\preccurlyeq_{s_1}$ and $\preccurlyeq_{s_2}$ could be considered alternatively. For instance the definitions previously stated can easily be extended to the Spearman correlation coefficient $\rho_X(s_1,s_2)$ (see Appendix B), that is the linear correlation coefficient between the random variables $F_{s_1}(s_1(X))$ and $F_{s_2}(s_2(X))$, where F_{s_i} denotes the cdf of $s_i(X)$, $i\in\{1,\ 2\}$.*

3.3 Median Rankings

The method for aggregating rankings we consider here relies on the so-called *median procedure*, which belongs to the family of *metric aggregation procedures* (see Barthélémy and Montjardet 1981 for further details). Let $d(.,.)$ be some metric or dissimilarity measure on the set of rankings on a finite set $\mathcal{Z}$. By definition, a *median ranking* among a profile $\Pi=\{\preceq_k\colon\ 1\leq k\leq K\}$ with respect to d is any ranking $\preceq_{med}$ on $\mathcal{Z}$ that minimizes the sum $\Delta_\Pi(\preceq)\overset{def}{=}\sum_{k=1}^K d(\preceq,\preceq_k)$ over the set $\mathbf{R}(\mathcal{Z})$ of all rankings $\preceq$ on $\mathcal{Z}$:

$$\Delta_\Pi(\preceq_{med})=\min_{\preceq\text{: ranking on }\mathcal{Z}}\Delta_\Pi(\preceq).$$

Notice that, when $\mathcal{Z}$ is of cardinality $N<\infty$, there are

$$\#\mathbf{R}(\mathcal{Z})=\sum_{k=1}^{N}(-1)^k\sum_{m=1}^{k}(-1)^m\left(\begin{array}{c}k\\m\end{array}\right)m^N$$

possible rankings on $\mathcal{Z}$ (that is the sum over k of the number of surjective mappings from $\{1,\dots,N\}$ to $\{1,\dots,k\}$) and in most cases, the computation of (metric) median rankings leads to NP-hard combinatorial optimization problems (see Wakabayashi 1998, Hudry 2004, Hudry 2008 and the references therein). It is worth noticing that a median ranking is far from being unique in general. One may immediately check for instance that any ranking among the profile made of all rankings on $\mathcal{Z} = \{1,\ 2\}$ is a median in Kendall sense, that is, for the metric d_τ. From a practical perspective, acceptably good solutions can be computed in a reasonable amount of time by means of metaheuristics such as simulated annealing, genetic algorithms or tabu search (see Spall 2003). The description of these computational aspects is beyond the scope of the present paper (see Charon and Hudry 1998 or Laguna et al. 1999 for instance). We also refer to recent work in Klementiev et al. (2009).

When it comes to preorders on a set $\mathcal{X}$ of infinite cardinality, defining a notion of aggregation becomes harder. Given a pseudo-metric such as d_τ and $B \geq 1$ scoring rules $s_1,\ \ldots,\ s_B$ on $\mathcal{X}$, the existence of $\bar{s}$ in $\mathcal{S}$ such that $\sum_{b=1}^{B} d_\tau(\bar{s}, s_b) = \min_s \sum_{b=1}^{B} d_\tau(s, s_b)$ is not guaranteed in general. However, when considering piecewise constant scoring rules with corresponding finite subpartition $\mathcal{P}$ of $\mathcal{X}$, the corresponding preorders are in one-to-one correspondence with rankings on $\mathcal{P}$ and the minimum distance is thus effectively attained.

Aggregation of piecewise constant scoring rules. Consider a finite collection of piecewise constant scoring rules $\Sigma_B = \{s_1,\ \ldots,\ s_B\}$ on $\mathcal{X}$, with $B \geq 1$.

Definition 14 (TRUE MEDIAN SCORING RULE). *Let $\mathcal{S}$ be a collection of scoring rules. We call $\bar{s}_B$ a* median scoring rule *for Σ_B with respect to $\mathcal{S}$ if*

$$\bar{s}_B = \arg\min_{s \in \mathcal{S}} \Delta_B(s),$$

where $\Delta_B(s) = \sum_{b=1}^{B} d_X(s, s_b)$ for $s \in \mathcal{S}$.

The empirical median scoring rule is obtained in a similar way, but the true distance d_X is replaced by its empirical counterpart $d_{\hat{\tau}_X}$, see Equation (3).

The ordinal approach. Metric aggregation procedures are not the only way to summarize a profile of rankings. The so-called "ordinal approach" provides a variety of alternative techniques for combining rankings (or, more generally, *preferences*), returning to the famous "Arrow's voting paradox". The ordinal approach consists of a series of duels (i.e., *pairwise comparisons*) as in Condorcet's method or successive tournaments as in the proportional voting Hare system, see Fishburn (1973). Such approaches have recently been the subject of a good deal of attention in the context of *preference learning* (also referred to as methods for *ranking by pairwise comparison*, see Hüllermeier et al. 2008 for instance).

Ranks vs. Rankings. Let $\Sigma_B = \{s_1,\ \ldots,\ s_B\}$, $B \geq 1$, be a collection of piecewise constant scoring rules and $\mathbf{X}'^{(m)} = \{X'_1, \ldots, X'_m\}$ a collection of $m \geq 1$ i.i.d. copies of the input variable X. When it comes to rank the observations X'_i "consensually", two strategies can be considered: (i) compute a "median ranking rule" based on the B rankings of the cells for the largest subpartition and use it for ranking the new data as previously described, or (ii) compute, for each scoring rule s_b, the related rank vector of the data set $\mathbf{X}'^{(m)}$, and then a "median rank vector", that is, a median ranking on the set $\mathbf{X}'^{(m)}$ (data lying in a same cell of the largest subpartition being tied). Although they are not equivalent, these two methods generally produce similar results, especially when m is large. Indeed, considering medians in the sense of probabilistic Kendall τ, it is sufficient to notice that the Kendall τ distance d_τ between rankings on $\mathbf{X}'^{(m)}$ induced by two piecewise constant rules s_1 and s_2 can be viewed as an empirical estimate of $d_X(s_1, s_2)$ based on the data set $\mathbf{X}'^{(m)}$. Now assume the collection Σ_B is obtained from training data $\mathcal{D}_n$. The difference between (i) and (ii) is that (i) does not use the data to be ranked $\mathbf{X}'^{(m)}$ but only relies on training data $\mathcal{D}_n$. However, when both the size of the training sample $\mathcal{D}_n$ and of the test data set $\mathbf{X}'^{(m)}$ are large, the two approaches lead to the optimization of related quantities.

4. Consistency of Aggregated Scoring Rules

We now provide statistical results for the aggregated scoring rules in the spirit of random forests (Breiman, 2001). In the context of classification, consistency theorems were derived in Biau et al. (2008). Conditions for consistency of piecewise constant scoring rules have been studied in Clémençon and Vayatis (2009c) and Clémençon et al. (2011). Here, we address the issue of AUC consistency of scoring rules obtained as medians over a profile of consistent randomized scoring rules for the (probabilistic) Kendall τ distance. A *randomized scoring rule* is a random element of the form $\widehat{s}_n(\cdot, Z)$, depending on both the training sample $\mathcal{D}_n = \{(X_1, Y_1),\ \dots,\ (X_n, Y_n)\}$ and a random variable Z, taking values over a measurable space $\mathcal{Z}$, independent of $\mathcal{D}_n$, which describes the randomization mechanism.

The AUC of a randomized scoring rule $\widehat{s}_n(\cdot, Z)$ is given by:

$$\begin{aligned}\mathrm{AUC}(\widehat{s}_n(\cdot,Z)) = \mathbb{P}\{\widehat{s}_n(X,Z) < \widehat{s}_n(X',Z) \mid (Y,Y') = (-1,+1)\} \\ + \frac{1}{2}\mathbb{P}\{\widehat{s}_n(X,Z) = \widehat{s}_n(X',Z) \mid (Y,Y') = (-1,+1)\},\end{aligned}$$

where the conditional probabilities are taken over the joint probability of independent copies (X,Y) and (X,Y') and Z, given the training data $\mathcal{D}_n$.

Definition 15 (AUC-CONSISTENCY) *The randomized scoring rule $\widehat{s}_n$ is said to be* AUC*-consistent (respectively, strongly* AUC*-consistent) when the convergence*

$$\mathrm{AUC}(\widehat{s}_n(\cdot,Z)) \to \mathrm{AUC}^* \textit{ as } \mathrm{n} \to \infty\ ,$$

holds in probability (respectively, almost-surely).

Let $B \geq 1$. Given $\mathcal{D}_n$, one may draw B i.i.d. copies $Z_1,\ \dots,\ Z_B$ of Z, yielding the collection $\widehat{\Sigma}_B$ of scoring rules $\widehat{s}_n(\cdot, Z_j)$, $1 \leq j \leq B$. Let $\mathcal{S}$ be a collection of scoring rules and suppose that $\bar{s}_B$ is a *median scoring rule* for the profile $\widehat{\Sigma}_B$ with respect to $\mathcal{S}$ in the sense of Definition 14. The next result shows that AUC-consistency is preserved for a median scoring rule of AUC-consistent randomized scoring rules.

Theorem 16 *(*CONSISTENCY AND AGGREGATION*) Set $B \geq 1$. Consider a class $\mathcal{S}$ of scoring rules. Assume that:*

- *the assumptions on the distribution of (X,Y) in Proposition 10 are fulfilled.*
- *the randomized scoring rule $\widehat{s}_n(\cdot,Z)$ is* AUC*-consistent (respectively, strongly* AUC*-consistent).*
- *for all $n, B \geq 1$, and for any sample $\mathcal{D}_n$, there exists a median scoring rule $\bar{s}_B \in \mathcal{S}$ for the collection $\{\widehat{s}_n(\cdot,Z_j), 1 \leq j \leq B\}$ with respect to $\mathcal{S}$.*

- *we have $\mathcal{S}^* \cap \mathcal{S} \neq \emptyset$.*

Then, the aggregated scoring rule $\bar{s}_B$ is AUC*-consistent (respectively, strongly* AUC*-consistent).*

We point out that the last assumption which states that the class $\mathcal{S}$ of candidate median scoring rules contains at least one optimal scoring function can be removed at the cost of an extra bias term in the rate bound. Consistency results are then derived by picking the median scoring rule, for each n, in a class $\mathcal{S}_n$ such that there exists a sequence $\tilde{s}_n \in \mathcal{S}_n$ which fulfills $\mathrm{AUC}(\tilde{\mathrm{s}}_\mathrm{n}) \to \mathrm{AUC}^*$ as $n \to \infty$. This remark covers the special case where $\widehat{s}_n(\cdot, Z)$ is a piecewise constant scoring rule with a range of cardinality $k_n \uparrow \infty$ and the median is taken over the set $\mathcal{S}_n$ of scoring functions with range of cardinality less than k_n^B. The bias is then of order $1/k_n^{2B}$ under mild smoothness conditions on ROC^*, as shown by Proposition 7 in Clémençon and Vayatis (2009b).

From a practical perspective, median computation is based on empirical versions of the probabilistic Kendall τ involved (see Equation (3)). The following result shows the existence of scoring rules that are asymptotically median with respect to d_X, provided that the class $\mathcal{S}$ over which the median is computed is not too complex. Here we formulate the result in terms of a VC major class of functions of finite dimension (see Dudley 1999 for instance). We first introduce the following notation, for any $s \in \mathcal{S}$:

$$\widehat{\Delta}_{B,m}(s) = \sum_{j=1}^{B} \widehat{d}_X(s, s_j) \ ,$$

where the estimate $\widehat{d_X}$ of d_X is based on $m \geq 1$ independent copies of X.

Theorem 17 *(*EMPIRICAL MEDIAN COMPUTATION*) Fix $B \geq 1$. Let $\Sigma_B = \{s_1, \ \ldots, \ s_B\}$ be a finite collection scoring rules and $\mathcal{S}$ a class of scoring rules which is a VC major class. We consider the empirical median scoring rule $\widetilde{s}_m = \arg\min_{s \in \mathcal{S}} \widehat{\Delta}_{B,m}(s)$. Then, as $m \to \infty$, we have*

$$\Delta_B(\widetilde{s}_m) \to \min_{s \in \mathcal{S}} \Delta_B(s) \ \textit{with probability one} \ .$$

The empirical aggregated scoring rule we consider in the next result relies on two data samples. The training sample $\mathcal{D}_n$, completed by the randomization on Z, leads to a collection of scoring rules which are instances of the randomized scoring rule. Then a sample $\mathbf{X'}^{(m)} = \{X'_1, \ldots, X'_m\}$ is used to compute the empirical median. Combining the two preceding theorems, we finally obtain the consistency result for the aggregated scoring rule.

Corollary 18 *Fix $B \geq 1$ and $\mathcal{S}$ a VC major class of scoring rules. Consider a training sample $\mathcal{D}_n$ of size n with i.i.d. copies of (X, Y) and a sample $\mathbf{X'}^{(m)}$ of size m with i.i.d. copies of X. We consider the collection $\widehat{\Sigma}_B$ of randomized scoring rules $\widehat{s}_n(\cdot, Z_j)$ in $\mathcal{S}$ built out of $\mathcal{D}_n$ and we introduce the empirical median of $\widehat{\Sigma}_B$ with respect to $\mathcal{S}$ obtained by using the test set $\mathbf{X'}^{(m)}$. We denote this fully empirical median scoring rule by $\widehat{s}_{n,m}$. If the assumptions of Theorem 16 are satisfied, then we have:*

$$\mathrm{AUC}(\widehat{s}_{n,m}) \xrightarrow{P} \mathrm{AUC}^* \ \textit{as} \ \mathrm{n}, \mathrm{m} \to \infty \ .$$

The results stated above can be extended to any median scoring rule based on a pseudo-metric d on the set of preorders on $\mathcal{S}$ which is equivalent to d_X, that is, such that $c_1 d_X \leq d \leq c_2 d_X$, with $0 < c_1 \leq c_2 < \infty$. Moreover, other complexity assumptions about the class $\mathcal{S}$ over which optimization is performed could be considered (see Clémençon et al. 2008). The present choice of VC major classes captures the complexity of scoring rules which will be considered in the next section (see Proposition 6 in Clémençon et al. 2011).

5. Ranking Forests

In this section, we introduce an implementation of the principles described in the previous sections for the aggregation of scoring rules. Here we focus on specific piecewise constant scoring rules based on ranking trees (Clémençon and Vayatis, 2009c; Clémençon et al., 2011). We propose various schemes for randomizing the features of these trees. We eventually describe the RANKING FOREST algorithm which extends to bipartite ranking the celebrated RANDOM FORESTS algorithm (Breiman, 1996; Amit and Geman, 1997; Breiman, 2001).

5.1 Tree-structured Scoring Rules

We consider piecewise constant scoring rules which can be represented in a left-right oriented binary tree. We recall that, in the context of classification, decision trees are very useful as they offer the possibility of interpretation for the selected classification rule. In the presence of classification data, one may entirely characterize a classification rule by means of a partition $\mathcal{P}$ of the input space $\mathcal{X}$ and a training set $\mathcal{D}_n = \{(X_i, Y_i) : 1 \leq i \leq n\}$ of i.i.d. copies of the pair (X, Y) through a *majority voting scheme*. Indeed, a new instance $x \in \mathcal{X}$ would receive the label corresponding to the most frequent one among the data points X_i within the cell $\mathcal{C} \in \mathcal{P}$ such that $x \in \mathcal{C}$. However, in bipartite ranking, the notion of local majority vote makes no sense since the ranking problem is of global nature. As a matter of fact, the issue is to rank the cells of the partition with respect to each other. It is assumed that ties among the ordered cells can be observed in the subsequent analysis and the usual MID-RANK convention is adopted. We refer to the Appendix A for a rigorous definition of the notion of *ranking* in the case of ties. We also point out that the term *partial ranking* is often used in this context (see Diaconis 1989, Fagin et al. 2006).

By restricting the search of candidates to the collection of piecewise constant scoring rules, the learning problem boils down here to finding a partition $\mathcal{P} = \{\mathcal{C}_k\}_{1 \leq k \leq K}$ of $\mathcal{X}$, with $1 \leq K < \infty$, together with a ranking $\preceq_{\mathcal{P}}$ of the $\mathcal{C}_k$'s (i.e., a preorder on $\mathcal{P}$), so that the ROC curve of the scoring rule given by

$$s_{\mathcal{P}, \preceq_{\mathcal{P}}}(x) = \sum_{k=1}^{K} (K - \mathcal{R}_{\preceq_{\mathcal{P}}}(\mathcal{C}_k) + 1) \cdot \mathbb{I}\{x \in \mathcal{C}_k\}$$

be as close as possible of ROC*, where $\mathcal{R}_{\preceq_{\mathcal{P}}}(\mathcal{C}_k)$ denotes the rank of $\mathcal{C}_k$, $1 \leq k \leq K$, among all cells of $\mathcal{P}$ according to $\preceq_{\mathcal{P}}$.

We now describe such scoring rules in the case where the partition arises from a tree structure. For such a partition, a ranking of the cells can be simply defined by equipping the tree with a left-right orientation. In order to describe how a ranking tree can be built so as to maximize AUC, further concepts are required. By *master ranking tree* $\mathcal{T}_D$, here we mean a complete, left-right oriented, rooted binary tree with depth $D \geq 1$. At depth $d = 0$, the entire input space $\mathcal{C}_{0,0} = \mathcal{X}$ forms its root. Every non terminal node (d, k), with $0 \leq d < D$ and $0 \leq k < 2^d$, is in correspondence with a subset $C_{d,k} \subset \mathcal{X}$, and has two siblings, each one corresponding to a subcell obtained by splitting $C_{d,k}$: the *left sibling* $C_{d+1,2k}$ is related to the leaf $(d+1, 2k)$, while the *right sibling* $C_{d+1,2k+1} = C_{d,k} \setminus C_{d+1,2k}$ is related to the leaf $(d+1, 2k+1)$ in the tree structure. We point out that an asymmetry is introduced at this point as the left sibling is assumed to have a lower rank (or higher score) than the right sibling in the ranking of the partition's cells. With this convention, it is easy to use any subtree $\mathcal{T} \subset \mathcal{T}_D$ as a ranking rule. A ranking of the terminal cells naturally results from the left-right orientation of the tree, the top of the list being represented by the cell in the bottom left corner of the tree, and is related to the scoring rule defined by: $\forall x \in \mathcal{X}$,

$$s_{\mathcal{T}}(x) = \sum_{(d,k):\ \text{terminal node of } \mathcal{T}} (2^D - 2^{D-d}k) \cdot \mathbb{I}\{x \in C_{d,k}\} \ .$$

The score value $s_{\mathcal{T}}(x)$ can be computed in a top-down manner, using the underlying "heap" structure. Starting from the initial value 2^D at the root node, at each subsequent inner node (d,k), $2^{D-(d+1)}$ is subtracted to the current value of the score if x moves down to the right sibling $(d+1, 2k+1)$, whereas one leaves the score unchanged if x moves down to the left sibling. The procedure is depicted in Figure 2.

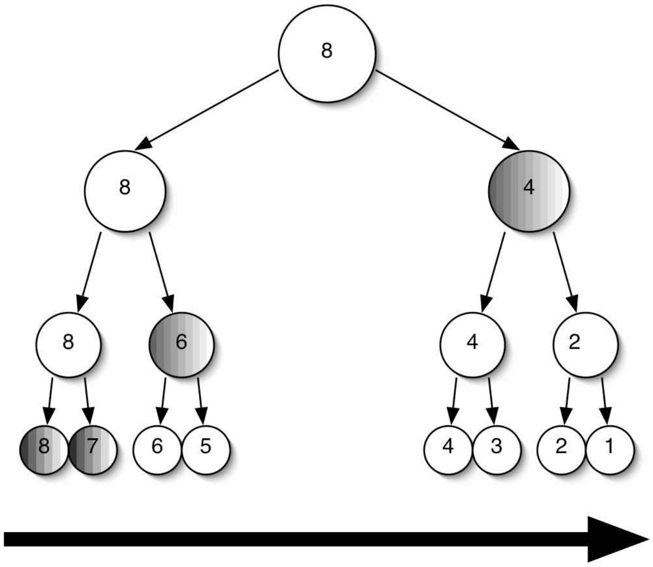


Figure 2: Ranking tree - the ranks can be read on the leaves of the tree from left (8 is the highest rank/score) to right (1 corresponds to the smallest rank/score). In case of a pruned tree (such as the one with leaves taken to be the shaded nodes), the orientation is conserved.

5.2 Feature Randomization in TREERANK

The concept of *bagging* (for **b**ootstrap **agg**regat**ing** technique) was introduced in Breiman (1996). The major novelty in the RANDOM FOREST method (Breiman, 2001) consisted in randomizing the features used for recursively splitting the nodes of the classification/regression trees involved in the committee-based prediction procedure. Our reference method for aggregating tree-based scoring rules is the TREERANK procedure (we refer to the Appendix and the papers Clémençon and Vayatis 2009c, Clémençon et al. 2011 for a full description). Beyond the specific structure of the master ranking tree, an additional ingredient in the growing stage is the splitting criterion. It turns out that a natural choice is a data-dependent and cost-sensitive classification error functional and its optimization can be performed with any binary classification method. This procedure for node splitting is called LEAFRANK. We point out that LEAFRANK implements a classifier and when this classifier is chosen to be a decision tree, this permits an additional randomization step. We thus propose two possible feature randomization schemes F_T for TREERANK and F_L for LEAFRANK.

F_T: *At each node (d,k) of the master ranking tree $\mathcal{T}_D$, draw at random a set of $q_0 \leq q$ indexes $\{i_1,\ldots,i_{q_0}\} \subset \{1,\ldots,q\}$. Implement the* LEAFRANK *splitting procedure based on the descriptor $(X^{(i_1)},\ldots,X^{(i_{q_0})})$ to split the cell $C_{d,k}$.*

F_L: *For each node (d,k) of the master ranking tree $\mathcal{T}_D$, at each node of the cost-sensitive classification tree describing the split of the cell $\mathcal{C}_{d,k}$ into two children, draw at random a set of $q_1 \leq q$ indexes $\{j_1,\ldots,j_{q_1}\} \subset \{1,\ldots,q\}$ and perform an axis-parallel cut using the descriptor $(X^{(j_1)},\ldots,X^{(j_{q_1})})$.*

We underline that, of course, the randomization strategy F_T can be applied to the TREERANK algorithm whatever the classification technique chosen for the splitting step. In addition, when the latter is itself a tree-based method, these randomization procedures do not exclude each other. At each node (d,k) of the ranking tree, one may first draw at random a collection $\mathcal{F}_{d,k}$ of q_0 features and then, when growing the cost-sensitive classification tree describing $C_{d,k}$'s split, divide each node based on a sub-collection of $q_1 \leq q_0$ features drawn at random among $\mathcal{F}_{d,k}$.

5.3 The RANKING FOREST Algorithm

Now that the rationale behind the RANKING FOREST procedure has been given, we describe its successive steps in detail. Based on a training sample $\mathcal{D}_n = \{(X_1,Y_1),\ldots,(X_n,Y_n)\}$, the algorithm is performed in three stages, as follows.

RANKING FOREST

1. **Parameters.** B number of bootstrap replicates, n^* bootstrap sample size, TREERANK tuning parameters (depth D and presence/absence of pruning), (F_T,F_L) feature randomization strategy, d pseudo-metric.

2. **Bootstrap profile makeup.**

   (a) (RESAMPLING STEP.) Build B independent bootstrap samples $\mathcal{D}^*_1$, $\ldots$, $\mathcal{D}^*_B$, by drawing with replacement $n^* \cdot B$ pairs among the original training sample $\mathcal{D}_n$.

   (b) (RANDOMIZED TREERANK.) For $b = 1$, $\ldots$, B, run TREERANK combined with the feature randomization method (F_T,F_L) based on the sample $\mathcal{D}^*_b$, yielding the ranking tree $\mathcal{T}^*_b$, related to the partition $\mathcal{P}^*_b$ of the space $\mathcal{X}$.

3. **Aggregation.** Compute the largest subpartition partition $\mathcal{P}^* = \bigcap_{b=1}^B \mathcal{P}^*_b$. Let $\preceq^*_b$ be the ranking of the cells of $\mathcal{P}^*$ induced by $\mathcal{T}^*_b$, $b = 1$, $\ldots$, B. Compute a median ranking $\preceq^*$ related to the bootstrap profile $\Pi^* = \{\preceq^*_b\colon\ 1 \leq b \leq B\}$ with respect to the metric d on $\mathbf{R}(\mathcal{P}^*)$:
$$\preceq^* = \underset{\preceq \in \mathbf{R}(\mathcal{P}^*)}{\arg\min}\, d_{\Pi^*}(\preceq),$$
as well as the scoring rule $s_{\preceq^*,\mathcal{P}^*}(\mathrm{x})$.

Remark 19 *(*ON TUNING PARAMETERS.*) As mentioned in 3.3, aggregating ranking rules is computationally expensive. The empirical results displayed in Section 6 suggest to aggregate several dozens of randomized ranking trees of moderate, or even small, depth built from bootstrap samples of size $n^* \leq n$.*

Remark 20 *(*"PLUG-IN" BAGGING.*) As pointed out in Clémençon and Vayatis (2009c) (see Remark 6 therein), given an ordered partition $(\mathcal{P}, \mathcal{R}_{\mathcal{P}})$ of the feature space $\mathcal{X}$, a "plug-in" estimate of the (optimal scoring) function $S = H_\eta \circ \eta$ can be automatically deduced from any ordered partition (or piecewise constant scoring rule equivalently) and the data $\mathcal{D}_n$, where H_η denotes the conditional cdf of $\eta(X)$ given $Y = -1$. This scoring rule is somehow canonical in the sense that, given $Y = -1$, $H(X)$ is distributed as a uniform r.v. on $[0,1]$, with H being the conditional distribution of X. Considering a partition $\mathcal{P} = \{\mathcal{C}_k\}_{1\leq k\leq K}$ equipped with a ranking $\mathcal{R}_{\mathcal{P}}$, the plug-in estimate is given by*

$$\widehat{S}_{\mathcal{P},\mathcal{R}_{\mathcal{P}}}(x) = \sum_{k=1}^{K} \widehat{\alpha}(R_k) \cdot \mathbb{I}\{x \in \mathcal{C}_k\}, \;\; x \in \mathcal{X},$$

where $R_k = \bigcup_{l:\, \mathcal{R}(k)\leq\mathcal{R}(l)} C_l$. Notice that, as a scoring rule, $\widehat{S}_{\mathcal{P},\mathcal{R}_{\mathcal{P}}}$ yields the same ranking as $s_{\mathcal{P},\mathcal{R}_{\mathcal{P}}}$, provided that $\widehat{\alpha}(\mathcal{C}_k) > 0$ for all $k = 1, \ldots, K$. Adapting the idea proposed in Section 6.1 of Breiman (1996) in the classification context, an alternative to the rank aggregation approach proposed here naturally consists in computing the average of the piecewise-constant scoring rules $\widetilde{S}^*_{\mathcal{T}^*_b}$ thus defined by the bootstrap ranking trees and consider the rankings induced by the latter. This method we call "plug-in bagging" is however less effective in many situations, due to the inaccuracy/variability of the probability estimates involved.

Ranking stability. Let $\Theta = \mathcal{X} \times \{-1, +1\}$. From the view developed in this paper, a ranking algorithm is a function $\mathbf{S}$ that maps any data sample $\mathcal{D}_n \in \Theta^n$, $n \geq 1$, to a scoring rule $\widehat{s}_n$. In the ranking context, we will say that a learning algorithm is "stable" when the preorder on $\mathcal{X}$ it outputs is not much affected by small changes in the training set. We propose a natural way of measuring ranking (in)stability, through the computation of the following quantity:

$$\mathbf{Instab}_n(\mathbf{S}) = \mathbb{E}\left(d_X\left(\widehat{s}_n, \widehat{s}'_n\right)\right) \;, \tag{4}$$

where the expectation is taken over two independent training samples $\mathcal{D}_n$ and $\mathcal{D}'_n$, both made of n i.i.d. copies of the pair (X,Y), and $\widehat{s}_n = \mathbf{S}(\mathcal{D}_n)$, $\widehat{s}'_n = \mathbf{S}(\mathcal{D}'_n)$. Incidentally, we highlight the fact that the bootstrap stage of RANKING FOREST can be used for assessing the stability of the base ranking algorithm. Indeed, set $\widehat{s}^{(b)}_{n^*} = \mathbf{S}(\mathcal{D}^*_b)$ and $\widehat{s}^{(b')}_{n^*} = \mathbf{S}(\mathcal{D}^*_{b'})$ obtained from two bootstrap samples. Then, the quantity:

$$\widehat{\mathbf{Instab}}_n(\mathbf{S}) = \frac{2}{B(B-1)} \sum_{1\leq b<b'\leq B} \widehat{d}_X\left(\widehat{s}^{(b)}_{n^*}, \widehat{s}^{(b')}_{n^*}\right) \;,$$

can be possibly interpreted as a bootstrap estimate of (4).

We finally underline that the outputs of the RANKING FOREST can also be used for monitoring ranking performance, in an analogous fashion to RANDOM FOREST in the classification/regression context (see Section 3.1 in Breiman 2001 and the references therein). An *out-of-bag* estimate of the AUC criterion can be obtained by considering, for all pairs (X,Y) and (X',Y') in the original training sample, those ranking trees that are built from bootstrap samples containing neither of them, avoiding this way the use of a test data set.

6. Numerical Experiments

The purpose of this section is to measure the impact of aggregation with resampling and feature randomization on the performance of the TREERANK/LEAFRANK procedure.

Data sets. We have considered artificial data sets where class-conditional distributions of X goven $Y = \pm 1$ are gaussian in dimensions 10 and 20. Three examples are considered here:

- *RF 10* - class-conditional distributions have the same means ($\mu_+ = \mu_- = 0$) but different covariance matrices ($\Sigma_+ = \mathbf{Id}_{10}$ and $\Sigma_- = 1.023 \cdot \mathbf{Id}_{10}$); optimal AUC is $\text{AUC}^* = 0.76$;
- *RF 20* - class-conditional distributions have different mean vectors ($||\mu_+ - \mu_-|| = 0.9$) and covariance matrices ($\Sigma_+ = \mathbf{Id}_{20}$ and $\Sigma_- = 1.23 \cdot \mathbf{Id}_{20}$); optimal AUC is $\text{AUC}^* = 0.77$;
- *RF 10 sparse* - class-conditional distributions have a 6-dimensional marginal distribution in common, and the regression function $\eta(x)$ depends on four components of the input vector X onlyoptimal AUC is $\text{AUC}^* = 0.89$.

With these data sets, the series of experiments below capture the influence on ranking performance of separability, dimension, and sparsity.

Sample sizes. In order to quantify the impact of bagging and random feature selection on the accuracy/stability of the resulting ranking rule, the algorithm has been run under various configurations for each data set on 30 independent training samples for each sample size ranging from $n = 250$ to $n = 3000$. The test sample was taken of size 3000 in all experiments.

Variants of TREERANK *and parameters.* In the intensive comparisons we have performed, we have considered the following variants:

- Plain TREERANK/LEAFRANK - in this version, all input dimensions are involved in the splitting stage; the maximum depth of the master ranking tree is 10, and the maximum depth of the ranking tree using orthogonal splits in the LEAFRANK procedure is 8 for the use case *RF 10 sparse* and also 10 for the two others.
- BAGGING RANKING TREES - the *bagging* version uses the plain TREERANK/LEAFRANK as described above with bootstrap samples of size $B = 20$, $B = 50$, and $B = 100$.
- RANKING FORESTS - the *forest* version involves additional parameters for feature randomization which can affect both the master ranking tree (F_T for TREERANK) and the splitting rule (F_L for LEAFRANK); these parameters indicate the number of dimensions randomly chosen along which the best split is chosen ; we have tried six different sets of parameters (Cases 1 to 6) where F_T takes values 3, 5, and 10 (or 20 for the data set *RF 20*), and F_L takes values 1, 3, and 5 (plus 10 for the data set *RF 20*); bootstrap samples are chosen of size $B = 1$ (single tree with feature randomization), $B = 20$, $B = 50$, and $B = 100$.

In the case of bagging and forests, aggregation is performed by taking the pointwise median value of ranks for the collection of ranking trees which have been estimated on each bootstrap sample. This choice allows for very fast evaluations of the aggregated scoring rule (see the last paragraph of Section 3.3 for a justification).

Performance. For each variant and each set of parameters and sample size, we performed 30 replications using independent training sets. These replications are used to derive performance results on a same test set. Performance is measured through a collection of indicators:

- $\overline{\text{AUC}}$ and $\widehat{\sigma}^2$ - Average AUC and standar type error are computed based on the test sample results over the 30 replications;

- ΔEnv - this indicator quantifies the accuracy of the variant through the relative improvement of the envelope on the ROC curve over the 30 replications compared to the plain TREERANK/LEAFRANK (e.g., if $\Delta\text{Env} = -30\%$ for BAGGING it means that the envelope of the ROC curve is 30% narrower than with TREERANK); the more negative the better the performance accuracy;

- **Instab**$_\tau$ - Instability measure applied to the ranking algorithm (e.g., Ranking Forest), estimate of (4), which reproduces the quantity $\widehat{\mathbf{Instab}}_n(\mathbf{S})$ using the Kendal τ as a distance; the smaller the quantity the more stable the method;

- DCG and AVE - the *Discounted Cumulative Gain* and the *Average Precision* provide measures which are sensitive to the top ranked instances; they can both be expressed as conditional linear rank statistics (see Clémençon and Vayatis 2007 and Clémençon and Vayatis 2009a) with score-generating function given by $1/(\ln(1+x))$ (DCG) or $1/x$ (AP);

- HR@u% - the *Hit Ratio* at u% is a relative count of positive instances among a proportion u of best scored instances.

These indicators capture the most important properties as far as quality assessment for scoring rules is concerned: average and local performance, stability of the rule, accuracy of ROC performance. *Results and comments.* Results are collected in a series of Tables 1, 2, 3, 4, 5, 6. We also report enveloppes on ROC curves over the series of replications of the experiments with the same parameters (see Figures 3 and 4). We study in particular the impact of mixed effects of randomization with sample size (Tables 1, 2, 3) or aggregation (Tables 4, 5, 6). Our main observations are the following:

- The sample size of the training set has a moderate impact on performance of RANKING FOREST while it helps significantly single trees in the plain TREERANK;

- In the case of small sample sizes, RANKING FOREST with little randomization (Cases 2 and 5) boost performance compared to the plain TREERANK;

- Increasing the amount of aggregation always improves performance and accuracy except in some situations in the non-sparse data sets (little randomization $F_T = d$, B large);

- BAGGING with $B = 20$ ranking trees already improves plain TREERANK dramatically;

- Randomization reveals its power in the sparse data set; when all input variables are relevant, highly randomized strategies (Cases 4 and 6) may fail to capture good scoring rules unless a large amount of ranking trees are aggregated (B above 50).

These empirical results aim at illustrating the effect of the combination of rank aggregation and random feature selection on ranking accuracy/stability. A complete and detailed empirical analysis of the merits and limitations of RANKING FOREST is beyond the scope of this paper and it will be the object of future work.

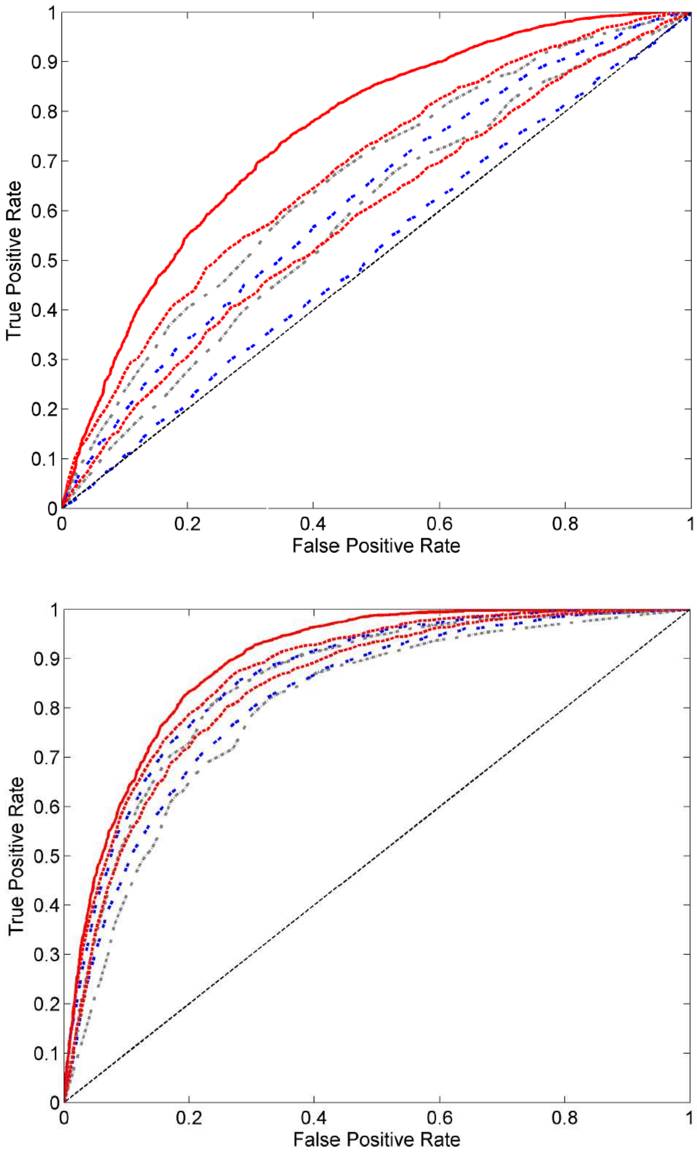
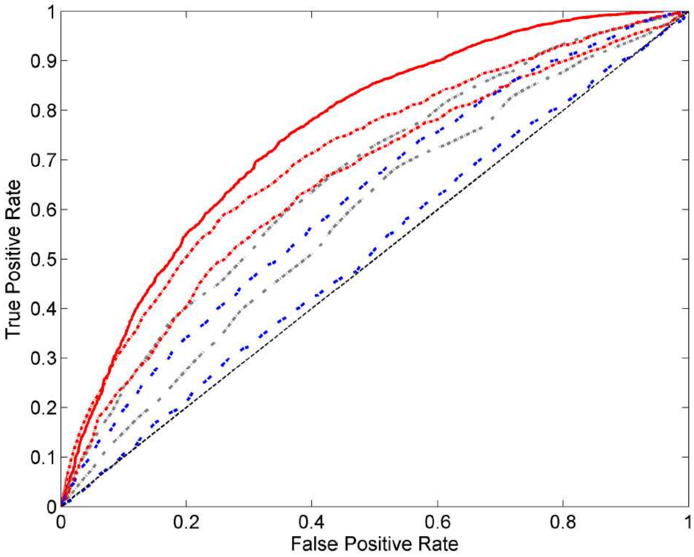


Figure 3: Comparison of envelopes on ROC curves - Results obtained with RANKING FORESTS with $B = 50$ (blue, double dashed) and 100 (red, solid, dashed). The upper display shows results on the data set *RF 10* while the lower display corresponds to the curves obtained on the data set *RF 10 sparse*. RANKING FORESTS used correspond to Case 3, training size is 2000, and optimal ROC curve is in thick red.

7. Conclusion

The major contribution of the paper was to show how to apply the principles of the RANDOM FOREST approach to the ranking/scoring task. Several ways of randomizing and aggregating ranking



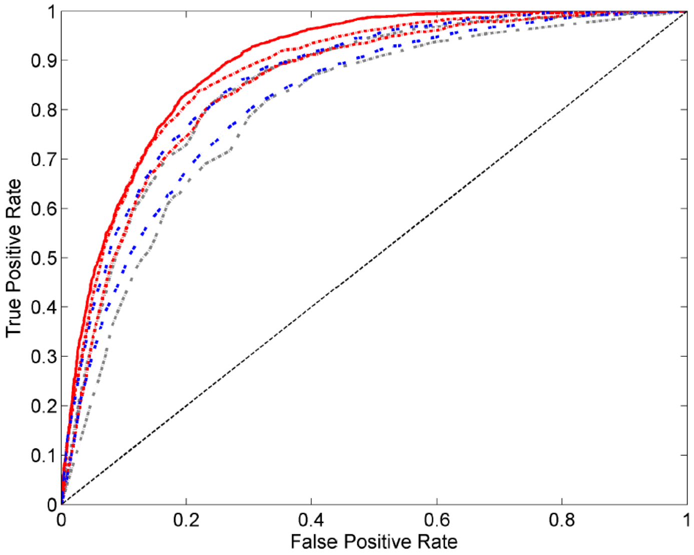


Figure 4: Comparison of envelopes on ROC curves - Results obtained with BAGGING (red, solid and dashed) and RANKING FORESTS (blue, double dashed) with $B = 50$. The upper display shows results on the data set *RF 10* while the lower display corresponds to the curves obtained on the data set *RF 10 sparse*. RANKING FORESTS used correspond to Case 3, training size is 2000, and optimal ROC curve is in thick red.

trees, such as those produced by the TREERANK algorithm, have been rigorously described. We proposed a specific notion of *stability* in the ranking setup and provided some preliminary back-

RF 10 - AUC* $= 0.756$ - dependence on aggregation

	F_T	F_L	B	$\overline{\text{AUC}}$ $(\pm\widehat{\sigma})$	ΔEnv	**Instab**$_\tau$	DCG	AVE	HR @10%	HR @20%
TreeRank	-	-	-	0.628 (± 0.013)	-	0.013	1.574	0.59	66%	64%
Bagging	-	-	20	0.678 (± 0.010)	−25%	0.010	1.708	0.64	77%	74%
			50	0.686 (± 0.008)	−29%	0.009	1.745	0.64	78%	74%
			100	0.689 (± 0.009)	−29%	0.009	1.819	0.65	78%	74%
Forest Case 1	5	5	1	0.508 (± 0.027)	+65%	0.016	1.563	0.50	49%	50%
			20	0.550 (± 0.026)	+55%	0.015	2.059	0.53	57%	55%
			50	0.567 (± 0.025)	+46%	0.015	2.210	0.55	59%	57%
			100	0.642 (± 0.016)	−7%	0.011	2.288	0.61	71%	67%
Forest Case No. 2	10	5	1	0.525 (± 0.025)	+68%	0.015	1.564	0.51	52%	52%
			20	0.577 (± 0.024)	+22%	0.014	2.012	0.56	61%	59%
			50	0.615 (± 0.020)	+21%	0.013	2.187	0.58	67%	64%
			100	0.585 (± 0.025)	+34%	0.014	2.357	0.56	62%	60%
Forest Case 3	5	3	1	0.512 (± 0.024)	+61%	0.016	1.564	0.50	49%	49%
			20	0.546 (± 0.024)	+35%	0.015	2.047	0.53	56%	54%
			50	0.577 (± 0.025)	+35%	0.014	2.215	0.56	61%	59%
			100	0.648 (± 0.019)	+23%	0.011	2.294	0.61	72%	68%
Forest Case 4	3	3	1	0.512 (± 0.023)	+51%	0.015	1.570	0.50	47%	49%
			20	0.537 (± 0.026)	+27%	0.015	2.067	0.52	54%	53%
			50	0.563 (± 0.028)	+42%	0.015	2.249	0.54	58%	57%
			100	0.595 (± 0.019)	0%	0.014	2.345	0.57	64%	61%
Forest Case 5	10	3	1	0.516 (± 0.029)	+95%	0.016	1.564	0.51	51%	51%
			20	0.582 (± 0.022)	+32%	0.014	2.016	0.56	62%	59%
			50	0.616 (± 0.022)	+11%	0.013	2.161	0.59	67%	64%
			100	0.579 (± 0.023)	+30%	0.014	2.423	0.56	61%	59%
Forest Case 6	3	1	1	0.517 (± 0.028)	+81%	0.016	1.567	0.51	51%	52%
			20	0.545 (± 0.026)	+38%	0.015	2.075	0.53	56%	55%
			50	0.565 (± 0.024)	+28%	0.015	2.224	0.55	59%	57%
			100	0.647 (± 0.016)	+3%	0.011	2.306	0.61	70%	67%

Table 1: Comparison of TREERANK/LEAFRANK and BAGGING with RANKING FORESTS - Impact of randomization (F_T, F_L) and resampling with aggregation (B) on the data set *RF 10* with training sample size $n = 2000$.

ground theory for ranking rule aggregation. Encouraging experimental results based on artificial data have also been obtained, demonstrating how bagging combined with feature randomization may significantly enhance ranking accuracy and stability both at the same time. Truth be told, theoretical explanations for the success of RANKING FOREST in these situations are left to be found. Results obtained by Friedman and Hall (2007) or Grandvalet (2004) for the bagging approach in the classification/regression context suggest possible lines of research in this regard. At the same time, further experiments, based on real data sets in particular, will be carried out in a dedicated article in order to determine precisely the situations in which RANKING FOREST is competitive compared to alternative ranking methods.

RF 20 - AUC* $= 0.773$ - dependence on aggregation

	F_T	F_L	B	$\overline{\text{AUC}}$ $(\pm\widehat{\sigma})$	ΔEnv	**Instab$_\tau$**	DCG	AVE	HR @10%	HR @20%
TreeRank	-	-	-	0.613 (± 0.013)	-	0.013	1.614	0.59	67%	64%
Bagging	-	-	20	0.691 (± 0.009)	-32%	0.009	1.715	0.66	80%	75%
			50	0.699 (± 0.006)	-43%	0.008	1.816	0.66	81%	76%
Forest Case 1	10	10	1	0.534 (± 0.033)	$+120\%$	0.015	1.599	0.53	56%	56%
			20	0.623 (± 0.028)	$+78\%$	0.013	2.017	0.60	68%	65%
			50	0.667 (± 0.021)	$+33\%$	0.011	2.017	0.63	73%	70%
			100	0.726 (± 0.011)	-25%	0.007	2.160	0.67	80%	77%
Forest Case 2	20	10	1	0.551 (± 0.033)	$+114\%$	0.015	1.599	0.54	58%	57%
			20	0.673 (± 0.019)	$+28\%$	0.011	1.989	0.64	73%	70%
			50	0.706 (± 0.012)	-15%	0.009	2.104	0.66	77%	74%
			100	0.693 (± 0.014)	0%	0.009	2.250	0.65	76%	73%
Forest Case 3	10	5	1	0.534 (± 0.030)	$+100\%$	0.015	1.599	0.53	56%	55%
			20	0.625 (± 0.025)	$+64\%$	0.013	2.077	0.60	68%	65%
			50	0.675 (± 0.013)	-6%	0.011	2.179	0.64	75%	71%
			100	0.726 (± 0.009)	-35%	0.007	2.171	0.67	80%	77%
Forest Case 4	5	5	1	0.516 (± 0.038)	$+138\%$	0.016	1.599	0.52	53%	53%
			20	0.585 (± 0.030)	$+93\%$	0.014	2.050	0.57	63%	61%
			50	0.625 (± 0.026)	$+50\%$	0.013	2.217	0.60	67%	65%
			100	0.702 (± 0.013)	-16%	0.009	2.247	0.66	78%	74%
Forest Case 5	20	5	1	0.547 (± 0.034)	$+123\%$	0.015	1.598	0.54	58%	56%
			20	0.666 (± 0.020)	$+25\%$	0.011	2.007	0.63	74%	70%
			50	0.705 (± 0.011)	-23%	0.009	2.128	0.66	78%	74%
			100	0.658 (± 0.021)	$+24\%$	0.011	2.329	0.62	71%	69%
Forest Case 6	5	1	1	0.510 (± 0.040)	$+157\%$	0.016	1.597	0.51	52%	52%
			20	0.574 (± 0.035)	$+97\%$	0.015	2.120	0.56	61%	59%
			50	0.614 (± 0.027)	$+64\%$	0.014	2.238	0.59	67%	64%
			100	0.710 (± 0.011)	-19%	0.009	2.261	0.66	78%	75%

Table 2: Comparison of TREERANK/LEAFRANK and BAGGING with RANKING FORESTS - Impact of randomization (F_T, F_L) and resampling with aggregation (B) on the data set *RF 20* with training sample size $n = 2000$.

Appendix A. Axioms for Ranking Rules

Throughout this paper, we call a *ranking* of the elements of a set $\mathcal{Z}$ any *total preorder* on $\mathcal{Z}$, that is, a binary relation $\preceq$ for which the following axioms are checked.

1. (TOTALITY) For all $(z_1, z_2) \in \mathcal{Z}^2$, either $z_1 \preceq z_2$ or else $z_2 \preceq z_1$ holds.

2. (TRANSITIVITY) For all (z_1, z_2, z_3): if $z_1 \preceq z_2$ and $z_2 \preceq z_3$, then $z_1 \preceq z_3$.

When the assertions $z_1 \preceq z_2$ and $z_2 \preceq z_1$ hold both at the same time, we write $z_1 \asymp z_2$ and $z_1 \prec z_2$ when solely the first one is true. Assuming in addition that $\mathcal{Z}$ has finite cardinality $\#\mathcal{Z} < \infty$, the rank

RF 10 sparse - AUC* $= 0.89$ - dependence on aggregation B

	F_T	F_L	B	$\overline{\text{AUC}}$ $(\pm\widehat{\sigma})$	ΔEnv	**Instab**$_\tau$	DCG	AVE	HR @10%	HR @20%
TreeRank	-	-	-	0.826 (±0.007)	-	0.007	1.622	0.70	84%	83%
Bagging	-	-	20	0.865 (±0.004)	−30%	0.004	1.643	0.74	89%	88%
			50	0.867 (±0.003)	−35%	0.004	1.650	0.74	89%	88%
			100	0.868 (±0.003)	−36%	0.004	1.708	0.74	89%	88%
Forest Case 1	5	5	1	0.630 (±0.071)	+502%	0.014	1.632	0.58	66%	63%
			20	0.814 (±0.018)	+61%	0.008	1.977	0.71	86%	84%
			50	0.832 (±0.012)	+22%	0.006	2.163	0.72	88%	85%
			100	0.858 (±0.006)	−30%	0.004	2.110	0.74	90%	88%
Forest Case 2	10	5	1	0.636 (±0.083)	+588%	0.014	1.598	0.59	71%	66%
			20	0.845 (±0.010)	−12%	0.005	1.893	0.73	89%	86%
			50	0.863 (±0.005)	−43%	0.004	1.918	0.74	90%	88%
			100	0.869 (±0.003)	−51%	0.003	1.956	0.74	91%	89%
Forest Case 3	5	3	1	0.622 (±0.071)	+553%	0.014	1.607	0.57	64%	60%
			20	0.809 (±0.010)	+72%	0.008	2.060	0.71	86%	83%
			50	0.844 (±0.009)	−15%	0.005	2.089	0.73	89%	87%
			100	0.859 (±0.005)	−38%	0.004	2.133	0.74	90%	88%
Forest Case 4	3	3	1	0.580 (±0.083)	+672%	0.015	1.612	0.55	61%	59%
			20	0.772 (±0.036)	+195%	0.010	2.056	0.68	83%	79%
			50	0.829 (±0.015)	+39%	0.007	2.211	0.72	88%	85%
			100	0.849 (±0.008)	−10%	0.005	2.253	0.73	90%	87%
Forest Case 5	10	3	1	0.661 (±0.069)	+480%	0.014	1.602	0.60	69%	66%
			20	0.840 (±0.010)	−9%	0.006	1.926	0.73	88%	86%
			50	0.863 (±0.005)	−41%	0.004	1.974	0.74	90%	88%
			100	0.868 (±0.010)	−54%	0.003	1.990	0.74	91%	89%
Forest Case 6	3	1	1	0.593 (±0.073)	+566%	0.015	1.611	0.55	63%	60%
			20	0.745 (±0.036)	+228%	0.011	2.162	0.66	79%	76%
			50	0.807 (±0.026)	+108%	0.008	2.252	0.70	86%	83%
			100	0.835 (±0.010)	−6%	0.006	2.318	0.72	88%	85%

Table 3: Comparison of TREERANK/LEAFRANK and BAGGING with RANKING FORESTS - Impact of randomization (F_T, F_L) and resampling with aggregation (B) on the data set *RF 10 sparse* with training sample size $n = 2000$.

of any element $z \in \mathcal{Z}$ is given by

$$\mathcal{R}_{\preceq}(z) = \sum_{z' \in \mathcal{Z}} \left\{ \mathbb{I}\{z' \prec z\} + \frac{1}{2}\mathbb{I}\{z' \asymp z\} \right\},$$

when using the standard MID-RANK convention (Kendall, 1945), that is, by assigning to tied elements the average of the ranks they cover.

Any scoring rule $s : \mathcal{Z} \to \mathbb{R}$ naturally defines a ranking $\preceq_s$ on $\mathcal{Z}$: $\forall (z_1, z_2) \in \mathcal{Z}^2$, $z_1 \preceq_s z_2$ iff $s(z_1) \leq s(z_2)$. Equipped with these notations, it is clear that $\preceq_{\mathcal{R}_{\preceq}}$ coincides with $\preceq$ for any ranking $\preceq$ on a finite set $\mathcal{Z}$.

RF 10 - AUC$^* = 0.76$ - dependence on sample size

	F_T	F_L	n	$\overline{\text{AUC}}$ $(\pm\widehat{\sigma})$	ΔEnv	**Instab**$_\tau$	DCG	AVE	HR @10%	HR @20%
TreeRank	-	-	250	0.573 (± 0.024)	-	0.014	1.673	0.54	60%	58%
			500	0.576 (± 0.018)	-	0.014	1.607	0.54	59%	58%
			1000	0.595 (± 0.018)	-	0.014	1.583	0.56	62%	60%
			2000	0.628 (± 0.013)	-	0.013	1.574	0.59	66%	64%
			3000	0.632 (± 0.011)	-	0.013	1.560	0.59	66%	65%
Bagging	-	-	2000	0.678 (± 0.010)	-25%	0.010	1.708	0.64	77%	74%
			3000	0.678 (± 0.010)	-25%	0.010	1.708	0.64	77%	74%
Case 1	5	5	250	0.546 (± 0.023)	-16%	0.015	2.034	0.53	56%	54%
			500	0.544 (± 0.028)	$+40\%$	0.015	2.032	0.53	56%	55%
			1000	0.547 (± 0.026)	$+10\%$	0.015	2.009	0.53	57%	55%
			2000	0.550 (± 0.026)	$+55\%$	0.015	2.059	0.53	57%	55%
			3000	0.549 (± 0.019)	$+33\%$	0.015	2.034	0.53	55%	55%
Case 2	10	5	250	0.571 (± 0.025)	$+11\%$	0.015	1.990	0.55	60%	58%
			500	0.571 (± 0.030)	$+34\%$	0.015	1.984	0.55	60%	59%
			1000	0.578 (± 0.028)	$+41\%$	0.014	1.999	0.56	60%	59%
			2000	0.577 (± 0.024)	$+22\%$	0.014	2.012	0.56	61%	59%
			3000	0.585 (± 0.028)	$+76\%$	0.014	1.998	0.56	62%	60%
Case 3	5	3	250	0.546 (± 0.031)	$+7\%$	0.015	2.049	0.53	56%	55%
			500	0.556 (± 0.029)	$+40\%$	0.015	1.993	0.54	58%	57%
			1000	0.563 (± 0.023)	$+8\%$	0.015	2.024	0.54	58%	57%
			2000	0.546 (± 0.024)	$+35\%$	0.015	2.047	0.53	56%	54%
			3000	0.549 (± 0.019)	$+30\%$	0.015	2.026	0.53	56%	55%
Case 4	3	3	250	0.546 (± 0.023)	$+15\%$	0.015	2.090	0.53	55%	55%
			500	0.536 (± 0.028)	$+36\%$	0.015	2.071	0.52	54%	53%
			1000	0.540 (± 0.027)	$+15\%$	0.015	2.075	0.53	55%	54%
			2000	0.537 (± 0.026)	$+27\%$	0.015	2.067	0.52	54%	53%
			3000	0.536 (± 0.022)	$+55\%$	0.015	2.063	0.52	54%	54%
Case 5	10	3	250	0.588 (± 0.027)	$+5\%$	0.014	1.984	0.56	62%	60%
			500	0.570 (± 0.030)	$+65\%$	0.015	1.970	0.55	59%	58%
			1000	0.587 (± 0.023)	$+16\%$	0.014	1.971	0.56	63%	60%
			2000	0.582 (± 0.022)	$+32\%$	0.014	2.016	0.56	62%	59%
			3000	0.587 (± 0.026)	$+83\%$	0.014	1.991	0.57	63%	60%
Case 6	3	1	250	0.546 (± 0.028)	$+8\%$	0.015	2.085	0.53	56%	55%
			500	0.543 (± 0.024)	$+11\%$	0.015	2.077	0.53	55%	54%
			1000	0.549 (± 0.026)	$+13\%$	0.015	2.066	0.53	56%	55%
			2000	0.545 (± 0.026)	$+38\%$	0.015	2.075	0.53	56%	55%
			3000	0.546 (± 0.026)	$+71\%$	0.015	2.065	0.53	56%	55%

Table 4: Comparison of TREERANK/LEAFRANK and BAGGING with RANKING FORESTS - Impact of randomization (F_T, F_L) and resampling with sample size (n) on the data set *RF 10* for $B = 20$.

Appendix B. Agreement Between Rankings

The most widely used approach to the *rank aggregation* issue relies on the concept of *measure of agreement* between rankings which uses *pseudo-metrics*. Since the seminal contribution of Kemeny

RF 20 - AUC$^* = 0.77$ - dependence on sample size

	F_T	F_L	n	$\overline{\text{AUC}}$ $(\pm\widehat{\sigma})$	ΔEnv	**Instab**$_\tau$	DCG	AVE	HR @10%	HR @20%
TreeRank	-	-	250	0.561 (±0.019)	-	0.014	1.742	0.55	57%	58%
			500	0.579 (±0.018)	-	0.014	1.666	0.56	60%	59%
			1000	0.593 (±0.014)	-	0.014	1.626	0.57	63%	62%
			2000	0.613 (±0.013)	-	0.013	1.614	0.59	67%	65%
			3000	0.621 (±0.013)	-	0.013	1.597	0.59	67%	65%
Bagging	-	-	2000	0.691 (±0.009)	−32%	0.009	1.715	0.66	80%	75%
			3000	0.691 (±0.009)	−32%	0.009	1.715	0.66	80%	75%
Case 1	10	10	250	0.612 (±0.026)	+25%	0.014	2.019	0.59	67%	64%
			500	0.630 (±0.029)	+41%	0.013	2.018	0.61	69%	66%
			1000	0.628 (±0.025)	+44%	0.013	2.024	0.60	68%	66%
			2000	0.623 (±0.028)	+78%	0.013	2.017	0.60	68%	65%
			3000	0.636 (±0.029)	+54%	0.012	2.012	0.61	67%	65%
Case 2	20	10	250	0.646 (±0.027)	+27%	0.012	1.964	0.62	71%	68%
			500	0.660 (±0.018)	+6%	0.012	1.945	0.63	72%	69%
			1000	0.666 (±0.019)	+23%	0.011	1.984	0.63	73%	70%
			2000	0.673 (±0.019)	+28%	0.011	1.989	0.64	73%	70%
			3000	0.665 (±0.017)	+17%	0.011	1.997	0.63	73%	70%
Case 3	10	5	250	0.610 (±0.030)	+69%	0.014	2.039	0.59	66%	63%
			500	0.617 (±0.033)	+56%	0.013	2.027	0.59	66%	64%
			1000	0.621 (±0.024)	+44%	0.013	2.035	0.60	67%	65%
			2000	0.625 (±0.025)	+64%	0.013	2.077	0.60	68%	65%
			3000	0.631 (±0.025)	+55%	0.013	2.039	0.61	69%	66%
Case 4	5	5	250	0.568 (±0.036)	+82%	0.015	2.088	0.56	61%	59%
			500	0.579 (±0.018)	+47%	0.014	2.064	0.58	63%	61%
			1000	0.585 (±0.041)	+155%	0.014	2.060	0.57	63%	61%
			2000	0.585 (±0.030)	+93%	0.014	2.050	0.57	63%	61%
			3000	0.585 (±0.030)	+88%	0.014	2.052	0.57	63%	61%
Case 5	20	5	250	0.631 (±0.018)	−4%	0.013	1.962	0.61	69%	67%
			500	0.658 (±0.021)	+4%	0.012	1.941	0.62	72%	69%
			1000	0.659 (±0.022)	+25%	0.012	1.988	0.63	72%	69%
			2000	0.666 (±0.020)	+25%	0.011	2.007	0.63	74%	70%
			3000	0.670 (±0.021)	+46%	0.011	1.978	0.63	73%	70%
Case 6	5	1	250	0.561 (±0.033)	+57%	0.015	2.099	0.55	59%	57%
			500	0.570 (±0.028)	+32%	0.015	2.061	0.56	61%	59%
			1000	0.571 (±0.031)	+119%	0.015	2.066	0.56	60%	59%
			2000	0.574 (±0.035)	+97%	0.015	2.120	0.56	61%	59%
			3000	0.570 (±0.032)	+88%	0.015	2.053	0.56	61%	60%

Table 5: Comparison of TreeRank/LeafRank and Bagging with Ranking Forests - Impact of randomization (F_T, F_L) and resampling with sample size (n) on the data set *RF 10* for $B = 20$.

(1959), numerous ways of measuring agreement have been proposed in the literature. Here we re-

RF 10 sparse - AUC* $= 0.89$ - dependence on sample size

	F_T	F_L	n	$\overline{\text{AUC}}$ $(\pm\widehat{\sigma})$	ΔEnv	**Instab**$_\tau$	DCG	AVE	HR @10%	HR @20%
TreeRank	-	-	250	0.749 (±0.022)	-	0.010	1.739	0.63	74%	74%
			500	0.771 (±0.015)	-	0.008	1.662	0.65	76%	76%
			1000	0.806 (±0.009)	-	0.008	1.637	0.68	80%	80%
			2000	0.827 (±0.007)	-	0.007	1.622	0.70	84%	83%
			3000	0.836 (±0.006)	-	0.007	1.602	0.70	85%	84%
Bagging	-	-	2000	0.865 (±0.004)	−30%	0.004	1.643	0.74	89%	88%
			3000	0.865 (±0.004)	−30%	0.004	1.643	0.74	89%	88%
Case 1	5	5	250	0.808 (±0.020)	−28%	0.008	2.010	0.71	87%	36%
			500	0.814 (±0.024)	+32%	0.008	1.958	0.71	86%	83%
			1000	0.862 (±0.005)	−49%	0.004	1.701	0.74	89%	88%
			2000	0.814 (±0.018)	+61%	0.008	1.977	0.71	86%	84%
			3000	0.870 (±0.005)	−19%	0.004	1.670	0.74	90%	88%
Case 2	10	5	250	0.835 (±0.012)	−57%	0.006	1.869	0.72	89%	86%
			500	0.841 (±0.011)	−36%	0.006	1.839	0.73	89%	86%
			1000	0.845 (±0.009)	−30%	0.006	1.853	0.73	90%	86%
			2000	0.845 (±0.010)	−12%	0.005	1.893	0.73	89%	86%
			3000	0.848 (±0.011)	+12%	0.006	1.851	0.73	89%	86%
Case 3	5	3	250	0.795 (±0.027)	−13%	0.009	2.014	0.70	86%	82%
			500	0.810 (±0.023)	+17%	0.008	1.984	0.71	86%	83%
			1000	0.811 (±0.020)	+40%	0.008	1.966	0.71	86%	83%
			2000	0.809 (±0.020)	+72%	0.008	2.060	0.71	86%	83%
			3000	0.809 (±0.023)	+110%	0.008	1.979	0.70	86%	83%
Case 4	3	3	250	0.764 (±0.042)	+27%	0.010	2.114	0.68	82%	78%
			500	0.773 (±0.038)	+115%	0.010	2.068	0.68	83%	79%
			1000	0.780 (±0.031)	+105%	0.009	2.063	0.69	83%	80%
			2000	0.772 (±0.036)	+195%	0.010	2.056	0.68	83%	79%
			3000	0.783 (±0.036)	+280%	0.009	2.044	0.69	83%	80%
Case 5	10	3	250	0.828 (±0.016)	−48%	0.007	1.931	0.72	87%	85%
			500	0.836 (±0.014)	−21%	0.006	1.883	0.72	88%	86%
			1000	0.841 (±0.012)	−9%	0.006	1.876	0.73	89%	86%
			2000	0.840 (±0.010)	+9%	0.006	1.926	0.73	88%	86%
			3000	0.843 (±0.008)	+5%	0.006	1.893	0.73	89%	86%
Case 6	3	1	250	0.724 (±0.049)	+32%	0.012	2.149	0.65	77%	74%
			500	0.757 (±0.035)	+76%	0.011	2.085	0.67	81%	78%
			1000	0.742 (±0.045)	+198%	0.011	2.096	0.66	79%	76%
			2000	0.745 (±0.036)	+228%	0.011	2.162	0.66	79%	76%
			3000	0.728 (±0.049)	+350%	0.012	2.079	0.65	78%	75%

Table 6: Comparison of TREERANK/LEAFRANK and BAGGING with RANKING FORESTS - Impact of randomization (F_T, F_L) and resampling with sample size (n) on the data set *RF 10 sparse* for $B = 20$.

view three popular choices, originally introduced in the context of *nonparametric statistical testing* (see Fagin et al. 2003 for instance).

Let $\preceq$ and $\preceq'$ be two rankings on a finite set $\mathcal{Z} = \{z_1, \ldots, z_K\}$. The notation $\mathcal{R}_{\preceq}(z)$ is used for the rank of the element z according to the ranking $\preceq$.
Kendall τ. Consider the quantity $d_\tau(\preceq, \preceq')$, obtained by summing up all the terms

$$\begin{aligned}U_{i,j}(\preceq, \preceq') = \mathbb{I}\{(\mathcal{R}_{\preceq}(z_i) - \mathcal{R}_{\preceq}(z_j))(\mathcal{R}_{\preceq'}(z_i) - \mathcal{R}_{\preceq'}(z_j)) < 0\}& \\ + \frac{1}{2}\mathbb{I}\{\mathcal{R}_{\preceq}(z_i) = s_{\preceq}(z_j),\ \mathcal{R}_{\preceq'}(z_i) \neq \mathcal{R}_{\preceq'}(z_j)\}& \\ + \frac{1}{2}\mathbb{I}\{\mathcal{R}_{\preceq'}(z_i) = \mathcal{R}_{\preceq'}(z_j),\ \mathcal{R}_{\preceq}(z_i) \neq \mathcal{R}_{\preceq}(z_j)\}&\end{aligned}$$

over all pairs (z_i, z_j) such that $1 \leq i < j \leq K$. It counts, among the $K(K-1)$ pairs of $\mathcal{Z}$'s elements, how many are "discording", assigning the weight $1/2$ when two elements are tied in one ranking but not in the other. The Kendall τ is obtained by renormalizing this distance:

$$\tau(\preceq, \preceq') = 1 - \frac{4}{K(K-1)} d_\tau(\preceq, \preceq').$$

Large values of $\tau(\preceq, \preceq')$ indicate agreement (or similarity) between $\preceq$ and $\preceq'$: it ranges from -1 (full disagreement) to 1 (full agreement). It is worth noticing that it can be computed in $O((K \log K)/\log\log K)$ time (see Bansal and Fernandez-Baca 2009).
Spearman footrule. Another natural distance between rankings is defined by considering the l_1-metric between the corresponding rank vectors:

$$d_1(\preceq,\ \preceq') = \sum_{i=1}^{K} |\mathcal{R}_{\preceq}(z_i) - \mathcal{R}_{\preceq'}(z_i)|.$$

The affine transformation given by

$$F(\preceq, \preceq') = 1 - \frac{3}{K^2 - 1} d_1(\preceq,\ \preceq').$$

is known as the Spearman footrule measure of agreement and takes its values in $[-1, +1]$.
Spearman rank-order correlation. Considering instead the l_2-metric

$$d_2(\preceq,\ \preceq') = \sum_{i=1}^{K} (\mathcal{R}_{\preceq}(z_i) - \mathcal{R}_{\preceq'}(z_i))^2$$

leads to the Spearman ρ coefficient:

$$\rho(\preceq,\ \preceq') = 1 - \frac{6}{K(K^2-1)} d_2(\preceq,\ \preceq').$$

Remark 21 *(*EQUIVALENCE.*) It should be noticed that these three measures of agreement are equivalent in the sense that:*

$$\begin{aligned} c_1\left(1 - \rho(\preceq,\ \preceq')\right) \leq&\quad (1 - F(\preceq,\ \preceq'))^2 \leq \quad c_2\left(1 - \rho(\preceq,\ \preceq')\right), \\ c_3\left(1 - \tau(\preceq,\ \preceq')\right) \leq&\quad 1 - F(\preceq,\ \preceq') \leq \quad c_4\left(1 - \tau(\preceq,\ \preceq')\right), \end{aligned}$$

with $c_2 = K^2/(2(K^2-1)) = Kc_1$ and $c_4 = 3K/(2(K+1)) = 2c_3$; see Theorem 13 in Fagin et al. (2006).

We point out that many fashions of measuring agreement or distance between rankings have been considered in the literature, see Mielke and Berry (2001). Well-known alternatives to the measures recalled above are the Cayley/Kemeny distance (Kemeny, 1959) and variants for top k-lists (Fagin et al., 2006), in order to focus on the "best instances" (see Clémençon and Vayatis 2007). Many other distances between rankings could naturally be deduced through suitable extensions of *word metrics* on the symmetric groups on finite sets (see Howie 2000 or Deza and Deza 2009).

Appendix C. The TREERANK Algorithm

Here we briefly review the TREERANK method, on which the procedure we call RANKING FOREST crucially relies. One may refer to Clémençon and Vayatis (2009c) and Clémençon et al. (2011) for further details as well as rigorous statistical foundations for the algorithm. As for most tree-based techniques, a greedy top-down recursive partitioning stage based on a training sample $\mathcal{D}_n = \{(X_i, Y_i) : 1 \leq i \leq n\}$ is followed by a pruning procedure, where children of a same parent node are recursively merged until an estimate of the AUC performance criterion is maximized. A package for R statistical software (see http://www.r-project.com) implementing TREERANK is available at http://treerank.sourceforge.net (see Baskiotis et al. 2009).

C.1 Growing Stage

The goal is here to grow a master ranking tree of large depth $D \geq 1$ with empirical AUC as large as possible. In order to describe this first stage, we introduce the following quantities. Let $\mathcal{C} \subset \mathcal{X}$, consider the empirical rate of negative (respectively, positive) instances lying in the region $\mathcal{C}$:

$$\widehat{\alpha}(\mathcal{C}) = \frac{1}{n}\sum_{i=1}^{n}\mathbb{I}\{X_i \in \mathcal{C},\, Y_i = -1\} \text{ and } \widehat{\beta}(\mathcal{C}) = \frac{1}{n}\sum_{i=1}^{n}\mathbb{I}\{X_i \in \mathcal{C},\, Y_i = +1\},$$

as well as $n(\mathcal{C}) = n(\widehat{\alpha}(\mathcal{C}) + \widehat{\beta}(\mathcal{C}))$ the number of data falling in $\mathcal{C}$.

One starts from the trivial partition $\mathcal{P}_0 = \{\mathcal{X}\}$ at root node $(0,0)$ (we set $\mathcal{C}_{0,0} = \mathcal{X}$) and proceeds recursively as follows. A tree-structured scoring rule $s(x)$ described by an oriented tree, with outer leaves forming a partition $\mathcal{P}$ of the input space, is refined by splitting a cell $\mathcal{C} \in \mathcal{P}$ into two subcells: $\mathcal{C}'$ denoting the left child and $\mathcal{C}'' = \mathcal{C} \setminus \mathcal{C}'$ the right one. Let $s'(x)$ be the scoring rule thus obtained. From the perspective of AUC maximization, one seeks a subregion $\mathcal{C}'$ maximizing the gain $\Delta_{\widehat{\mathrm{AUC}}}(\mathcal{C}, \mathcal{C}')$ in terms of empirical AUC induced by the split, which may be written as:

$$\widehat{\mathrm{AUC}}(s') - \widehat{\mathrm{AUC}}(s) = \frac{1}{2}\{\widehat{\alpha}(\mathcal{C})\widehat{\beta}(\mathcal{C}') - \widehat{\beta}(\mathcal{C})\widehat{\alpha}(\mathcal{C}')\}.$$

Therefore, taking the rate of positive instances within the cell $\mathcal{C}$, $\widehat{p}(\mathcal{C}) = \widehat{\alpha}(\mathcal{C}) \cdot n/n(\mathcal{C})$ namely, as cost for the type I error (i.e., predicting label $+1$ when $Y = -1$) and $1 - \widehat{p}(\mathcal{C})$ as cost for the type II error, the quantity $1 - \Delta_{\widehat{\mathrm{AUC}}}(\mathcal{C}, \mathcal{C}')$ may be viewed as the *cost-sensitive empirical misclassification error* of the classifier $C(X) = 2 \cdot \mathbb{I}\{X \in \mathcal{C}'\} - 1$ on $\mathcal{C}$ up to a multiplicative factor, $4\widehat{p}(\mathcal{C})(1 - \widehat{p}(\mathcal{C}))$ precisely. Hence, once the local cost $\widehat{p}(\mathcal{C})$ is computed, any binary classification method can be straightforwardly adapted in order to perform the splitting step. Here, splits are obtained using empirical-cost sensitive versions of the standard CART algorithm with axis-parallel splits, this one-step procedure for AUC maximization being called LEAFRANK in Clémençon et al. (2011). As depicted by Figure 5, the growing stage appears as a recursive implementation of a cost-sensitive

CART procedure with a cost updated at each node of the ranking tree, equal to the local rate of positive instances within the node to split, see Section 3 of Clémençon et al. (2011).

C.2 Pruning Stage

The way the master ranking tree $\mathcal{T}_D$ obtained at the end of the growing stage is pruned is entirely similar to the one described in Breiman et al. (1984), the sole difference lying in the fact that here, for any $\lambda > 0$, one seeks a subtree $\mathcal{T} \subset \mathcal{T}_D$ that maximizes the penalized empirical AUC

$$\widehat{\text{AUC}}(s_{\mathcal{T}}) - \lambda \cdot |\mathcal{T}|,$$

where $|\mathcal{T}|$ denotes the number of terminal leaves of $\mathcal{T}$, the constant being next picked using N-fold cross validation.

The fact that alternative complexity-based penalization procedures, inspired from recent nonparametric model selection methods and leading to the concept of *structural* AUC *maximization*, can be successfully used for pruning ranking trees has also been pointed up in Section 4.2 of Clémençon et al. (2011). However, the resampling-based technique previously mentioned is preferred to such pruning schemes in practice, insofar as it does not require, in contrast, to specify any tuning constant. Following in the footsteps of Arlot (2009) in the classification setup, estimation of the ideal penalty through bootstrap methods could arise as the answer to this issue. This question is beyond the scope of the present paper but will soon be tackled.

C.3 Some Practical Considerations

Like other types of decision trees, ranking trees (based on perpendicular splits) have a number of crucial advantages. Concerning interpretability first, it should be noticed that they produce ranking rules that can be easily visualized through the binary tree graphic, see Figure 5, the rank/score of an instance $x \in \mathcal{X}$ being obtained through checking of a nested combination of simple rules of the form "$X^{(k)} \geq t$" or "$X^{(k)} < t$". In addition, ranking trees can adapt straightforwardly to situations where some data are missing and/or some predictor variables are categorical and some monitoring tools helping to evaluate the relative importance of each predictor variable $X^{(k)}$ or to depict the partial dependence of the prediction rule on a subset of input variables are readily available. These facets are described in section 5 of Clémençon et al. (2011). From a computational perspective now, we also underline that the tree structure makes the computation of consensus rankings much more tractable, we refer to Appendix D for further details.

Appendix D. On Computing the Largest Subpartition

We now briefly explain how to make crucial use of the fact that the partitions of $\mathcal{X}$ we consider here are tree-structured to compute the largest subpartition they induce. Let $\mathcal{P}_1 = \{\mathcal{C}_k^{(1)}\}_{1 \leq k \leq K_1}$ and $\mathcal{P}_2 = \{\mathcal{C}_k^{(2)}\}_{1 \leq k \leq K_2}$ be two partitions of $\mathcal{X}$, related to (ranking) trees $\mathcal{T}_1$ and $\mathcal{T}_2$ respectively. For any $k \in \{1, \dots, K_1\}$, the collection of subsets of the form $\mathcal{C}_k^{(1)} \cap \mathcal{C}_l^{(2)}$, $1 \leq l \leq K_2$, can be obtained by extending the $\mathcal{T}_1$ tree structure the following way. At the $\mathcal{T}_1$'s terminal leave defining the cell $\mathcal{C}_k^{(1)}$, add a subtree corresponding to $\mathcal{T}_2$ with root $\mathcal{C}_k^{(1)}$: the terminal nodes of the resulting subtree, starting at the global root $\mathcal{X}$, correspond to the desired collection of subsets (notice that some of these can be empty), see Figure 6 below. Of course, this scheme can be iterated in order to recover

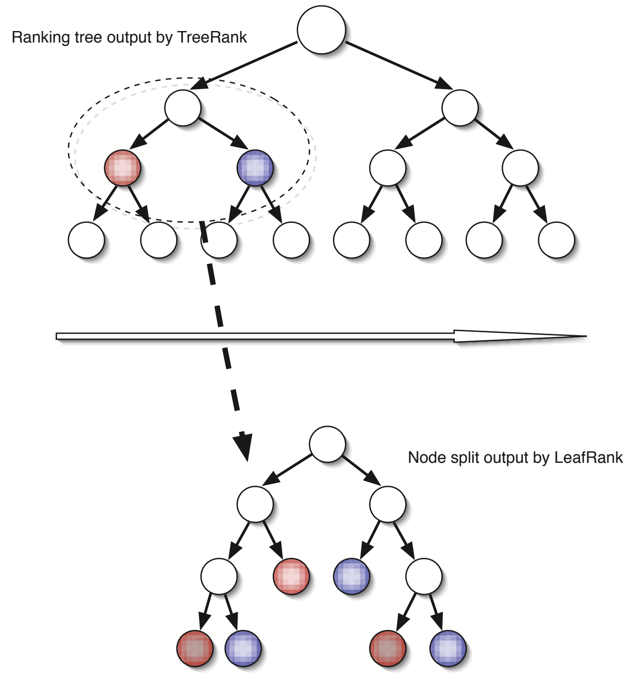

Ranking tree output by TreeRank
Node split output by LeafRank

Figure 5: The TREERANK algorithm as a recursive implementation of cost-sensitive CART.

all the cells of the subpartition induced by $B > 2$ tree-structured partitions. For obvious reasons of computational nature, one should start with the most complex tree and bind progressively less and less complex trees as one goes along.

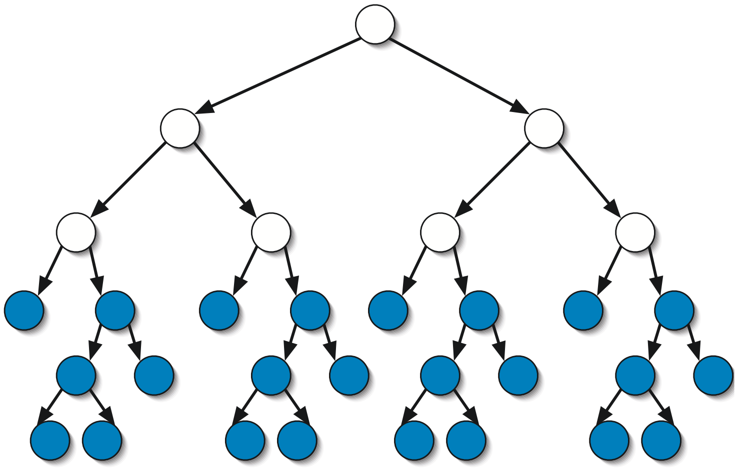


Figure 6: Characterizing the largest subpartition induced by tree-structured partitions.

Appendix E. Proofs

This section contains the proofs of the theoretical results presented in the core of the paper.

E.1 Proof of Proposition 9

Recall that $\tau_X(s_1,s_2)=1-2d_X(s_1,s_2)$, where $d_X(s_1,s_2)$ is given by:

$$\mathbb{P}\{(s_1(X)-s_1(X'))\cdot(s_2(X)-s_2(X'))<0\}+\frac{1}{2}\mathbb{P}\{s_1(X)=s_1(x'),\ s_2(X)\neq s_2(X')\}$$
$$+\frac{1}{2}\mathbb{P}\{s_1(X)\neq s_1(x'),\ s_2(X)=s_2(X')\}.$$

Observe first that, for all s, AUC(s) may be written as:

$$\mathbb{P}\{(s(X)-s(X'))\cdot(Y-Y')>0\}/(2p(1-p))+\mathbb{P}\{s(X)=s(X'),\ Y\neq Y'\}/(4p(1-p)).$$

Notice also that, using Jensen's inequality, one easily obtain that $2p(1-p)|\mathrm{AUC}(s_1)-\mathrm{AUC}(s_2)|$ is bounded by the expectation of the random variable

$$\mathbb{I}\{(s_1(X)-s_1(X'))\cdot(s_2(X)-s_2(X'))>0\}+\frac{1}{2}\mathbb{I}\{s_1(X)=s_1(X')\}\cdot\mathbb{I}\{s_2(X)\neq s_2(X')\}+$$
$$\frac{1}{2}\mathbb{I}\{s_1(X)\neq s_1(X')\}\cdot\mathbb{I}\{s_2(X)=s_2(X')\},$$

which is equal to $d_X(s_1,s_2)=(1-\tau_X(s_1,s_2))/2$.

E.2 Proof of Proposition 10

Recall first that, for all $s\in\mathcal{S}$, the AUC deficit $2p(1-p)\{\mathrm{AUC}^*-\mathrm{AUC}(s)\}$ may be written as

$$\mathbb{E}\left[|\eta(X)-\eta(X')|\cdot\mathbb{I}\{(X,X')\in\Gamma_s\}\right]+\mathbb{P}\{s(X)=s(X'),\ (Y,Y')=(-1,+1)\},$$

with

$$\Gamma_s=\{(x,x')\in\mathcal{X}^2:\ (s(x)-s(x'))\cdot(\eta(x)-\eta(x'))<0\},$$

refer to Example 1 in Clémençon et al. (2008) for instance. Now, Hölder inequality combined with noise condition (1) shows that $\mathbb{P}\{(X,X')\in\Gamma_s\}$ is bounded by

$$\left(\mathbb{E}\left[|\eta(X)-\eta(X')|\cdot\mathbb{I}\{(X,X')\in\Gamma_s\}\right]\right)^{a/(1+a)}\times c^{1/(1+a)}.$$

Therefore, we have for all $s^*\in\mathcal{S}^*$:

$$d_X(\preccurlyeq_s,\preccurlyeq_{s^*})=\mathbb{P}\{(X,X')\in\Gamma_s\}+\frac{1}{2}\mathbb{P}\{s(X)=s(X')\}.$$

Notice that $p(1-p)\mathbb{P}\{s(X)=s(X')\mid(Y,Y')=(-1,+1)\}$ can be rewritten as

$$\mathbb{E}[\mathbb{I}\{s(X)=s(X')\}\cdot\eta(X')(1-\eta(X))]=\frac{1}{2}\mathbb{E}[\mathbb{I}\{s(X)=s(X')\}\cdot(\eta(X')+\eta(X)-2\eta(X)\eta(X'))],$$

which term can be easily shown to be larger than $\frac{1}{2}\mathbb{E}[\mathbb{I}\{s(X)=s(X')\}\cdot|\eta(X')-\eta(X)|]$. Using the same argument as above, we obtain that $\mathbb{P}\{s(X)=s(X')\}$ is bounded by

$$\left(\mathbb{E}\left[|\eta(X)-\eta(X')|\cdot\mathbb{I}\{s(X)=s(X')\}\right]\right)^{a/(1+a)}\times c^{1/(1+a)}.$$

Combined withe the bound previously established, this leads to the desired result.

E.3 Proof of Theorem 16

By virtue of Proposition 9, we have:

$$\mathrm{AUC}^* - \mathrm{AUC}(\bar{\mathrm{s}}_{\mathrm{B}}) \leq \frac{\mathrm{d}_{\mathrm{X}}(\mathrm{s}^*, \bar{\mathrm{s}}_{\mathrm{B}})}{2\mathrm{p}(1-\mathrm{p})},$$

for any $s^* \in \mathcal{S}^*$. Using now triangular inequality applied to the distance d_X between preorders on $\mathcal{X}$, one gets

$$d_X(s^*, \bar{s}_B) \leq d_X(s^*, \widehat{s}_n(., Z_j)) + d_X(\widehat{s}_n(., Z_j), \bar{s}_B),$$

for all $j \in \{1, \ldots, B\}$. Averaging then over j and using the fact that, if one chooses s^* in $\mathcal{S}$,

$$\sum_{j=1}^{B} d_X(\widehat{s}_n(., Z_j), \bar{s}_B) \leq \sum_{j=1}^{B} d_X(\widehat{s}_n(., Z_j), s^*),$$

one obtains that

$$d_X(s^*, \bar{s}_B) \leq \frac{2}{B} \sum_{j=1}^{B} d_X(\widehat{s}_n(., Z_j), s^*).$$

The desired result finally follows from Proposition 10 combined with the consistency assumption of the randomized scoring rule.

Remark 22 *Observe that, in the case where $\mathcal{S}$ is allowed to depend on n and one only assumes the existence of $\tilde{s}_n^* \in \mathcal{S}_n$ such that* $\mathrm{AUC}(\tilde{\mathrm{s}}_{\mathrm{n}}^*) \to \mathrm{AUC}^*$ *as $n \to \infty$ (relaxing thus the assumption $\mathcal{S} \cap \mathcal{S}^* \neq \emptyset$), the argument above leads to*

$$\mathrm{AUC}^* - \mathrm{AUC}(\bar{\mathrm{s}}_{\mathrm{B}}) \leq \frac{1}{2\mathrm{p}(1-\mathrm{p})} \left\{ \frac{2}{\mathrm{B}} \sum_{\mathrm{j}=1}^{\mathrm{B}} \mathrm{d}_{\mathrm{X}}(\widehat{\mathrm{s}}_{\mathrm{n}}(., \mathrm{Z}_{\mathrm{j}}), \mathrm{s}^*) + \mathrm{d}_{\mathrm{X}}(\tilde{\mathrm{s}}_{\mathrm{n}}^*, \mathrm{s}^*) \right\}.$$

which shows that AUC *consistency of the median still holds true.*

E.4 Proof of Theorem 17

Observe that we have:

$$\begin{aligned} \Delta_B(\tilde{s}_m) - \min_{s \in \mathcal{S}} \Delta_B(s) &\leq 2 \cdot \sup_{s \in \mathcal{S}} |\widehat{\Delta}_{B,m}(s) - \Delta_B(s)| \\ &\leq 2 \sum_{j=1}^{B} \sup_{s \in \mathcal{S}} |\widehat{d}_X(s, s_j) - d_X(s, s_j)|. \end{aligned}$$

Now, it results from the strong Law of Large Numbers for U-processes stated in Corollary 5.2.3 in de la Pena and Giné (1999) that $\sup_{s \in \mathcal{S}} |\widehat{d}_X(s, s_j) - d_X(s, s_j)| \to 0$ as $N \to \infty$, for all $j = 1, \ldots, B$. The convergence rate $O_{\mathbb{P}}(m^{-1/2})$ follows from the Central Limit Theorem for U-processes given in Theorem 5.3.7 in de la Pena and Giné (1999).

E.5 Proof of Corollary 18

Reproducing the argument of Theorem 16, one gets:

$$d_X(s^*, \hat{s}_{n,m}) \leq \frac{1}{B}\sum_{j=1}^{B} d_X(\widehat{s}_n(.,Z_j), s^*) + \frac{1}{B}\sum_{j=1}^{B} d_X(\widehat{s}_n(.,Z_j), \hat{s}_{n,m}).$$

As in Theorem 17's proof, we also have:

$$\frac{1}{B}\sum_{j=1}^{B}\{d_X(\widehat{s}_n(.,Z_j),\widehat{s}_{n,m}) \quad - \quad d_X(\widehat{s}_n(.,Z_j),\bar{s}_B)\} \quad \leq \quad 2 \quad \cdot \quad \sup_{(s,s')\in\mathcal{S}^2} |\widehat{d_X}(s,s') \quad - \quad d_X(s,s')|.$$

Using again Corollary 5.2.3 in de la Pena and Giné (1999), we obtain that the term on the right hand side of the bound above vanishes as $m \to \infty$. Now the desired result immediately follows from Theorem 16.

References

S. Agarwal, T. Graepel, R. Herbrich, S. Har-Peled, and D. Roth. Generalization bounds for the area under the ROC curve. *J. Mach. Learn. Res.*, 6:393–425, 2005.

Y. Amit and D. Geman. Shape quantization and recognition with randomized trees. *Neural Computation*, 9:1545–1587, 1997.

S. Arlot. Model selection by resampling techniques. *Electronic Journal of Statistics*, 3:557–624, 2009.

M.S. Bansal and D. Fernandez-Baca. Computing distances between partial rankings. *Information Processing Letters*, 109:238–241, 2009.

J.P. Barthélémy and B. Montjardet. The median procedure in cluster analysis and social choice theory. *Mathematical Social Sciences*, 1:235–267, 1981.

N. Baskiotis, S. Clémençon, M. Depecker, and N. Vayatis. R-implementation of the TreeRank algorithm. *Submitted for publication*, 2009.

N. Betzler, M.R. Fellows, J. Guo, R. Niedermeier, and F.A. Rosamond. Computing kemeny rankings, parameterized by the average kt-distance. In *Proceedings of the 2nd International Workshop on Computational Social Choice*, 2008.

G. Biau, L. Devroye, and G. Lugosi. Consistency of Random Forests. *J. Mach. Learn. Res.*, 9: 2039–2057, 2008.

L. Breiman. Bagging predictors. *Machine Learning*, 26:123–140, 1996.

L. Breiman. Random forests. *Machine Learning*, 45(1):5–32, 2001.

L. Breiman, J. Friedman, R. Olshen, and C. Stone. *Classification and Regression Trees*. Wadsworth and Brooks, 1984.

I. Charon and O. Hudry. Lamarckian genetic algorithms applied to the aggregation of preferences. *Annals of Operations Research*, 80:281–297, 1998.

S. Clémençon and N. Vayatis. Ranking the best instances. *Journal of Machine Learning Research*, 8:2671–2699, 2007.

S. Clémençon and N. Vayatis. Empirical performance maximization based on linear rank statistics. In *NIPS*, volume 3559 of *Lecture Notes in Computer Science*, pages 1–15. Springer, 2009a.

S. Clémençon and N. Vayatis. Overlaying classifiers: a practical approach to optimal scoring. *To appear in Constructive Approximation*, 2009b.

S. Clémençon and N. Vayatis. Tree-based ranking methods. *IEEE Transactions on Information Theory*, 55(9):4316–4336, 2009c.

S. Clémençon, G. Lugosi, and N. Vayatis. Ranking and scoring using empirical risk minimization. In *Proceedings of COLT*, 2005.

S. Clémençon, G. Lugosi, and N. Vayatis. Ranking and empirical risk minimization of U-statistics. *The Annals of Statistics*, 36(2):844–874, 2008.

S. Clémençon, M. Depecker, and N. Vayatis. Bagging ranking trees. *Proceedings of ICMLA'09*, pages 658–663, 2009.

S. Clémençon, M. Depecker, and N. Vayatis. AUC-optimization and the two-sample problem. In *Proceedings of NIPS'09*, 2010.

S. Clémençon, M. Depecker, and N. Vayatis. Adaptive partitioning schemes for bipartite ranking. *Machine Learning*, 83(1):31–69, 2011.

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

V. de la Pena and E. Giné. *Decoupling: from Dependence to Independence*. Springer, 1999.

M.M. Deza and E. Deza. *Encyclopedia of Distances*. Springer, 2009.

P. Diaconis. A generalization of spectral analysis with application to ranked data. *The Annals of Statistics*, 17(3):949–979, 1989.

R.M. Dudley. *Uniform Central Limit Theorems*. Cambridge University Press, 1999.

C. Dwork, R. Kumar, M. Naor, and D. Sivakumar. Rank aggregation methods for the Web. In *Proceedings of the 10th International WWW Conference*, pages 613–622, 2001.

J.P. Egan. *Signal Detection Theory and ROC Analysis*. Academic Press, 1975.

R. Fagin, R. Kumar, M. Mahdian, D. Sivakumar, and E. Vee. Comparing and aggregating rankings with ties. In *Proceedings of the 12-th WWW Conference*, pages 366–375, 2003.

R. Fagin, R. Kumar, M. Mahdian, D. Sivakumar, and E. Vee. Comparing partial rankings. *SIAM J. Discrete Mathematics*, 20(3):628–648, 2006.

P. Fishburn. *The Theory of Social Choice*. University Press, Princeton, 1973.

M.A. Fligner and J.S. Verducci (Eds.). *Probability Models and Statistical Analyses for Ranking Data*. Springer, 1993.

Y. Freund, R. D. Iyer, R. E. Schapire, and Y. Singer. An efficient boosting algorithm for combining preferences. *Journal of Machine Learning Research*, 4:933–969, 2003.

J. Friedman and P. Hall. On bagging and non-linear estimation. *Journal of Statistical Planning and Inference*, 137(3):669–683, 2007.

Y. Grandvalet. Bagging equalizes influence. *Machine Learning*, 55:251–270, 2004.

T. Hastie and R. Tibshirani. *Generalized Additive Models*. Chapman and Hall/CRC, 1990.

J.M. Hilbe. *Logistic Regression Models*. Chapman and Hall/CRC, 2009.

J. Howie. Hyperbolic groups. *In Groups and Applications, edited by V. Metaftsis, Ekdoseis Ziti, Thessaloniki*, pages 137–160, 2000.

O. Hudry. Computation of median orders: complexity results. *Annales du LAMSADE: Vol. 3. Proceedings of the Workshop on Computer Science and Decision Theory, DIMACS*, 163:179–214, 2004.

O. Hudry. NP-hardness results for the aggregation of linear orders into median orders. *Ann. Oper. Res.*, 163:63–88, 2008.

E. Hüllermeier, J. Fürnkranz, W. Cheng, and K. Brinker. Label ranking by learning pairwise preferences. *Artificial Intelligence*, 172:1897–1917, 2008.

I. Ilyas, W. Aref, and A. Elmagarmid. Joining ranked inputs in practice. In *Proceedings of the 28th International Conference on Very Large Databases*, pages 950–961, 2002.

J. G. Kemeny. Mathematics without numbers. *Daedalus*, (88):571–591, 1959.

M.G. Kendall. The treatment of ties in ranking problems. *Biometrika*, (33):239–251, 1945.

A. Klementiev, D. Roth, K. Small, and I. Titov. Unsupervised rank aggregation with domain-specific expertise. In *IJCAI'09: Proceedings of the 21st International Joint Conference on Artifical Intelligence*, pages 1101–1106, San Francisco, CA, USA, 2009. Morgan Kaufmann Publishers Inc.

M. Laguna, R. Marti, and V. Campos. Intensification and diversification with elite tabu search solutions for the linear ordering problem. *Computers and Operations Research*, 26(12):1217–1230, 1999.

G. Lebanon and J. Lafferty. Conditional models on the ranking poset. In *Proceedings of NIPS'03*, 2003.

B. Mandhani and M. Meila. Tractable search for learning exponential models of rankings. In *Proceedings of AISTATS, Vol. 5 of JMLR:W&CP 5*, 2009.

M. Meila, K. Phadnis, A. Patterson, and J. Bilmes. Consensus ranking under the exponential model. In *Conference on Artificial Intelligence (UAI)*, pages 729–734, 2007.

P.W. Mielke and K.J. Berry. *Permutation Methods*. Springer, 2001.

A. Nemirovski. *Lectures on Probability Theory and Statistics, Ecole d'ete de Probabilities de Saint-Flour XXVIII - 1998*, chapter Topics in Non-Parametric Statistics. Number 1738 in Lecture Notes in Mathematics. Springer, 2000.

D.M. Pennock, E. Horvitz, and C.L. Giles. Social choice theory and recommender systems: analysis of the foundations of collaborative filtering. In *National Conference on Artificial Intelligence*, pages 729–734, 2000.

J.C. Spall. *Introduction to Stochastic Search and Optimization: Estimation, Simulation, and Control*. John Wiley & Sons, 2003.

Y. Wakabayashi. The complexity of computing medians of relations. *Resenhas*, 3(3):323–349, 1998.