The TreeRank Tournament Algorithm for Multipartite Ranking

Sylvain Robbiano, Stéphan Clémençon

Journal of Nonparametric Statistics
Vol. 00, No. 00, Month 2010, 1–21

The TreeRank Tournament Algorithm for Multipartite Ranking

Stéphan Clémençon[a*] and Sylvain Robbiano[b]

[a]*4 LTCI UMR Telecom ParisTech/CNRS No. 5141*
Telecom ParisTech 46, rue Barrault - 75634 Paris cedex 13, France;
[b]*CIMFAV, Universitad de Valparaíso, Pedro Montt 2421, Valparaíso, Chile*

(*v3.7 released September 2009*)

Whereas a variety of efficient learning algorithms have been recently proposed to perform bipartite ranking tasks, cast as M-estimation problems, when $K \geq 3$, no method for optimizing the ROC manifold, or criteria summarizing the latter such as its volume, the gold standard for assessing performance in K-partite ranking, have been introduced in the statistical learning literature yet. It is the main purpose of this paper to describe at length an efficient approach to recursive maximization of the ROC surface, extending the TreeRank methodology originally tailored for the bipartite situation (*i.e.* when $K = 2$). The main barrier arises from the fact that, in contrast to the bipartite case, the VUS criterion of any scoring rule taking $K \geq 3$ values cannot be interpreted as a *cost-sensitive* misclassification error and no method is readily available to perform the recursive optimization stage. The learning algorithm we propose, called TreeRank Tournament, breaks it and builds recursively an ordered partition of the feature space, defining a piecewise scoring function whose ROC manifold can be remarkably interpreted as a statistical version of an adaptive piecewise linear approximant of the optimal ROC manifold. Rate bounds in sup norm desccribing the generalization ability of the scoring rule thus built are established and numerical results illustrating the performance of the TreeRank Tournament approach, compared to that of natural competitors such as aggregation methods, are also displayed.

Keywords: Multipartite ranking, ROC surface, VUS optimization, recursive partitioning

AMS Subject Classification: 62G10 ; 62C99

*Corresponding author. Stéphan Clémençon Email: stephan.clemencon@telecom-paristech.fr - Tel: +33 1 45 81 78 07 - Fax: +33 1 45 81 71 58

ISSN: 1048-5252 print/ISSN 1029-0311 online

DOI: 10.1080/1048525YYxxxxxxxx
http://www.informaworld.com

1. Introduction

The multipartite ranking problem refers to the situation where an ordinal label Y, taking its values in $\{1, \ldots, K\}$ with $K \geq 2$ say, is assigned to any random observation X and the goal is to learn, based on a training sample composed of independent labelled observations, how to rank new data in the same order as their (temporarily hidden) labels. Though of great simplicity, this formulation covers a wide variety applications: design of diagnosis support tools in medicine, credit-risk screening in finance, *etc.* Motivated by this broad range of applied problems, multipartite ranking has been the subject of a good deal of attention these last few years in the machine-learning literature. In particular, it is much documented in the bipartite situation, where the learning problem can be cast as ROC *curve optimization* or AUC *maximization* and for which theoretical results and specific learning algorithms are available, see Clémençon and Vayatis (2009), Agarwal et al. (2005) or Clémençon and Vayatis (2010) and the references therein for instance. In the general K-partite situation, this learning problem has been generally tackled from the perspective of *pairwise classification* or *preference learning*, see Clémençon et al. (2008), Freund et al. (2003). Although the *empirical risk minimization* paradigm can be applied to it with statistical guarantees (see Rajaram and Agarwal (2005)), no algorithm dedicated to the optimization of the ROC manifold, the extension of the ROC curve in the multipartite context, has been proposed in the statistical learning literature yet. Indeed, recent approaches are based on reducing K-partite ranking to a collection of $K-1$ (or $K(K-1)/2$) bipartite ranking tasks, see Hüllermeier et al. (2008) or Clémençon et al. (2013).

It is the goal of this article to propose a multipartite ranking algorithm for recursive optimization of the ROC manifold, extending the TreeRank approach originally introduced in the bipartite setup, see Clémençon et al. (2011). As will be explained at length in the following, in contrast with the bipartite case, the recursive optimization stage cannot be formulated as a *cost-sensitive* classification problem, for which "off-the-shelf" methods can be used. In the algorithm we propose, called TreeRank Tournament, local optimization of the ROC surface is performed by comparing the performance of the possible updates obtained by implementing the optimization step of the TreeRank algorithm applied to the bipartite ranking subproblems of the multipartite problem considered. This method is shown to produce a piecewise constant ranking rule, whose ROC manifold can be viewed as a statistical counterpart of an adaptive piecewise linear approximant of the optimal ROC manifold. This point is worth paying attention to, insofar as piecewise linear interpolants of the optimal ROC manifold are not ROC manifolds in general, in contrast to the bipartite situation. Rate bounds are next established under adequate assumptions. Beyond statistical grounds, numerical results are provided in order to support the empirical performance of the TreeRank Tournament algorithm, compared to that of its competitors.

The paper is structured as follows. A rigorous formulation of the multipartite ranking problem is given in section 2 and basic concepts of ROC analysis are briefly recalled. Section 3 describes at length the multipartite ranking algorithm we propose, called TreeRank Tournament. An adaptive approximation scheme of the optimal ROC surface is next introduced and analyzed in section 4. In particular, bounds for the error in sup norm are proved. This result is the key to the study of the generalization ability of the TreeRank Tournament algorithm in section 5. This theoretical guarantee is completed by numerical experiments illustrating the performance of the approach promoted. Technical proofs are deferred to the Appendix section.

2. Background and Preliminaries

It is the purpose of this section to recall crucial notions inherent to the formulation of the multipartite ranking issue and to performance evaluation in this context.

2.1. *Multipartite Ranking*

We start off with describing the probabilistic framework and setting out the main notations of the paper. Let $K \geq 2$. Here and throughout, Y is a discrete random variable, taking its values in the finite *ordinal* set $\mathcal{Y} = \{1, \ldots, K\}$ and X is a random vector defined on the same probability space, which models some (hopefully useful) observation for predicting Y. Typically, the r.v. X is valued in a subset $\mathcal{X}$ of a high-dimensional euclidean space, $\mathbb{R}^d$ with $d \geq 1$ say. The distribution of the pair (X, Y) is characterized by the marginal distribution of X, $F(dx)$, and the posterior probabilities $(\eta_1(x), \ldots, \eta_K(x)) = (\mathbb{P}\{Y = 1 \mid X = x\}, \ldots, \mathbb{P}\{Y = K \mid X = x\})$, $x \in \mathcal{X}$. We also set $p_k = \mathbb{P}\{Y = k\}$ and denote by $F_k(dx)$ the conditional distribution of X given $Y = k$ for $k = 1, \ldots, K$. Roughly speaking, the goal of K-partite ranking is to order all elements of the set $\mathcal{X}$ through a (measurable) scoring function $s : \mathcal{X} \to \mathbb{R}$ transporting the natural order on the real line onto $\mathcal{X}$ (namely, $\forall (x_1, x_2) \in \mathcal{X}^2$: $x_1 \leq_s x_2 \Leftrightarrow s(x_1) \leq s(x_2)$), in a way that Y and $s(X)$ tend to increase or decrease together with largest probability. Clearly, when $K = 3$ for instance, such a problem would be completely meaningless if one could exhibit points x_1 and x_2 in $\mathcal{X}^2$ such that, when passing from x_1 to x_2 for instance, the quantity $(dF_3/dF_2)(x)$ increases, while the likelihood ratio $(dF_2/dF_1)(x)$ decreases. As discussed at length in Clémençon et al. (2013), except in the bipartite situation, the formulation of the multipartite ranking problem involves restrictive conditions on the class distributions.

Condition 2.1: Let $K \geq 2$. For any $1 \leq k \leq l < K$, for all $(x_1, x_2) \in \mathcal{X}^2$:

$$\frac{dF_{l+1}}{dF_l}(x_1) < \frac{dF_{l+1}}{dF_l}(x_2) \Rightarrow \frac{dF_{k+1}}{dF_k}(x_1) \leq \frac{dF_{k+1}}{dF_k}(x_2).$$

Observe that the condition above is void when $K = 2$. For $K \geq 3$, it guarantees the existence of scoring functions $s^*(x)$ that can be expressed as a strictly increasing transform of the ratio $dF_{k+1}/dF_k(x)$ for any $k \in \{1, \ldots, K-1\}$, see Proposition 1 in Clémençon et al. (2013). Such scoring functions naturally form the set $\mathcal{S}^*$ of optimal elements for the K-partite ranking problem related to the *monotone likelihood ratio* collection of class distributions $\{F_1, \ldots, F_K\}$. Incidentally, we also recall that the regression function $\eta(x) = \mathbb{E}[Y \mid X = x]$ belongs to the optimal set $\mathcal{S}^*$ (see Assertion (3) in Proposition 1 of Clémençon et al. (2013)). Provided that Condition 2.1 is fulfilled, the K-partite ranking task consists of building from training data a scoring function that "nearly" ranks data in the same order as the elements of $\mathcal{S}^*$. The concept of ROC surface/manifold described below permits precisely to quantify performance in this context.

2.2. ROC *Analysis*

For any measurable scoring function s, we denote by $F_{s,k}(t) = \mathbb{P}\{s(X) \leq t \mid Y = k\}$ the conditional cumulative distribution function(cdf in short) of the random variable $s(X)$ given $Y = k$, for $1 \leq k \leq K$. When $s = \eta$, we shall denote the previous functions by F_k^*.

The ROC graphic of s is the set of points

$$M_{\mathbf{t}} = \left(F_{s,1}(t_1) - F_{s,1}(t_0),\ \ldots,\ F_{s,K}(t_K) - F_{s,K}(t_{K-1})\right),$$

where $-\infty = t_0 < t_1 \leq \ldots \leq t_{K-1} < t_K = \infty$. We have $F_{s,K}(t_K) = 1$ and $F_{s,1}(t_0) = 0$. Denoting by $\mathbb{I}\{\mathcal{E}\}$ the indicator function of any event $\mathcal{E}$, observe also that the coordinates of the point $M_{\mathbf{t}}$ coincides with the diagonal entries of the *confusion matrix* of the classification rule defined by thresholding $s(X)$ at the levels t_k, $1 \leq k < K$:

$$C_{s,\mathbf{t}}(X) = \sum_{k=1}^{K} k \cdot \mathbb{I}\{t_{k-1} < s(X) \leq t_k\}.$$

We have indeed $\mathbb{P}\left\{C_{s,\mathbf{t}}(X) = k \mid Y = k\right\} = F_{s,k}(t_k) - F_{s,k}(t_{k-1})$ for all k in $\{1,\ \ldots,\ K\}$. By convention, all possible discontinuities (due to possible jumps of the distributions F_k) are connected by parts of affine hyperplanes. The ROC graphic is then a continuous manifold of dimension $K-1$, referred to as "ROC manifold". For notational simplicity, we mainly restrict our attention to the case $K = 3$. However, the subsequent analysis can be straightforwardly extended to the general K-partite situation. When $K = 2$, the ROC manifold is a curve, image by the transform $(\alpha,\ \beta) \in [0,1]^2 \mapsto (1-\alpha,\ \beta)$ of the graph of a nondecreasing *càd-làg*[1] mapping $\alpha \in [0,1] \mapsto \mathrm{ROC}_{1,\,2}(s,\ \alpha)$, defined by

$$\mathrm{ROC}_{1,\,2}(s,\ \alpha) = 1 - F_{s,2} \circ F_{s,1}^{-1}(1-\alpha)$$

at points α such that $F_{s,1} \circ F_{s,1}^{-1}(1-\alpha) = 1-\alpha$, denoting by $W^{-1}(u) = \inf\{t \in]-\infty,\ +\infty] :\ W(t) \geq u\}$, $u \in [0,1]$, the generalized inverse of any cdf $W(t)$ on $\mathbb{R}$. Equipped with these notations, the ROC surface can be then viewed as the graph of a function $(\alpha,\ \gamma) \in (0,1)^2 \mapsto \mathrm{ROC}_s(\alpha,\ \gamma)$, where

$$\mathrm{ROC}_s(\alpha,\ \gamma) = \left(F_{s,2} \circ F_{s,3}^{-1}(1-\gamma) - F_{s,2} \circ F_{s,1}^{-1}(\alpha)\right)_+,$$

at points $(\alpha,\ \gamma)$ such that $F_{s,1} \circ F_{s,1}^{-1}(\alpha) = \alpha$ and $F_{s,3} \circ F_{s,3}^{-1}(1-\gamma) = 1-\gamma$, with $u_+ = \max(u,\ 0)$ for any $u \in \mathbb{R}$. Notice incidentally that, with the convention above, the ROC surface of a piecewise constant scoring function is piecewise planar. As proved in Clémençon et al. (2013), the ROC surface of $\mathcal{S}^*$'s elements, ROC^* say, dominates everywhere that of any other scoring function s under Condition 2.1: $\forall (\alpha,\gamma) \in (0,1)^2$, $\mathrm{ROC}_s(\alpha,\gamma) \leq \mathrm{ROC}^*(\alpha,\gamma)$. Refer to Clémençon et al. (2013) for further details (see Theorem 1 therein), as well as a list of properties of ROC surfaces. In particular, recall that ROC^* is concave. In this case, just like the ROC curve in the bipartite situation, the criterion ROC_s provides a way of measuring ranking performance and induces a partial preorder on the set of scoring functions. A scoring function s_1 will be said better than another one s_2 when $\mathrm{ROC}_{s_1}(\alpha,\gamma) \geq \mathrm{ROC}_{s_2}(\alpha,\gamma)$ for all $(\alpha,\gamma) \in (0,1)^2$. This functional

[1]Recall that, by definition, a *càd-làg* function $h : [0,1] \to \mathbb{R}$ is such that $\lim_{s \to t,\ s<t} h(s) = h(t-) < \infty$ for all $t \in]0,1]$ and $\lim_{s \to t,\ s>t} h(s) = h(t)$ for all $t \in [0,1[$. Its completed graph is obtained by connecting the points $(t, h(t-))$ and $(t, h(t))$, when they are not equal, by a vertical line segment and thus forms a continuous curve.

criterion can also be summarized by the *Volume Under the* ROC *Surface*

$$\mathrm{VUS(s)} \stackrel{\mathrm{def}}{=} \int\int \mathrm{ROC_s}(\alpha,\gamma)\mathrm{d}\alpha\mathrm{d}\gamma.$$

It can be interpreted in a probabilistic fashion as the "rate of concording 3-tuples", through the formula (see Scurfield (1996)):

$$\begin{aligned}\mathrm{VUS(s)} = {} & \mathbb{P}\{s(X_1) < s(X_2) < s(X_3)\} + \frac{1}{2}\mathbb{P}\{s(X_1) = s(X_2) < s(X_3)\} \\ & + \frac{1}{2}\mathbb{P}\{s(X_1) < s(X_2) = s(X_3)\} + \frac{1}{6}\mathbb{P}\{s(X_1) = s(X_2) = s(X_3)\}, \qquad (1)\end{aligned}$$

where $X_1,\ X_2$ and X_3 denote independent r.v.'s defined on the same probability space with respective distributions F_1, F_2 and F_3. See Appendix A for the extension of formula (1) to the general multipartite setup. Statistical versions of the ROC surface and of the VUS criterion are obtained by replacing the class distributions by their empirical counterparts. Eq. (1) extends the well-known formula related to the *Area Under the* ROC *curve*:

$$\mathrm{AUC}_{1,2}(s) = \int \mathrm{ROC}_{1,\,2}(s,\ \alpha)d\alpha = \mathbb{P}\{s(X_1) < s(X_2)\} + \frac{1}{2}\mathbb{P}\{s(X_1) = s(X_2)\}.$$

We may now rephrase the ranking task in a quantitative manner. The goal is to build, from training data, a scoring function s whose ROC surface is "as close as possible" to ROC^*. In such a functional framework, various ways of measuring "closeness" can be considered of course. In particular, we focus here on the following important cases:

$$\begin{aligned} d_\infty(s,s^*) &= \sup_{(\alpha,\gamma)\in(0,1)^2} |\mathrm{ROC}^*(\alpha,\gamma) - \mathrm{ROC}_s(\alpha,\gamma)|, \\ d_1(s,s^*) &= \int\int |\mathrm{ROC}^*(\alpha,\gamma) - \mathrm{ROC}_s(\alpha,\gamma)|d\alpha d\gamma = \mathrm{VUS}^* - \mathrm{VUS(s)}, \end{aligned}$$

where $s^* \in \mathcal{S}^*$ and $\mathrm{VUS}^* \stackrel{\mathrm{def}}{=} \mathrm{VUS}(s^*)$. We point out that the quantities above do not represent distances between the scoring functions but distances between their ROC surfaces. Whereas minimization of $d_1(s,s^*)$ is clearly equivalent to maximization of VUS(s), observe also that minimization of $d_\infty(s,s^*)$ can hardly be cast as a M-estimation problem, no empirical counterpart of the criterion to optimize being available.

2.3. *Bipartite Ranking and the* TreeRank *Algorithm*

In Clémençon and Vayatis (2009) (see also Clémençon et al. (2011)), a bipartite ranking algorithm optimizing directly the ROC curve in a recursive manner, called TreeRank, has been proposed and thoroughly studied. It produces an oriented partition of the feature space $\mathcal{X}$ (defining thus a ranking, for which elements of a same cell being viewed as ties). The process is described by a left-to-right oriented binary tree structure, termed *ranking tree*, with fixed maximum depth $J \geq 0$. At depth $j \leq J$, there are 2^j nodes, indexed by (j,k) with $0 \leq k < 2^j$. The root node represents the whole feature space $\mathcal{C}_{0,0} = \mathcal{X}$ and each *internal node* (j,k) with $j < J$ and $k \in \{0,\ \ldots,\ 2^j - 1\}$ corresponds to

a subset $\mathcal{C}_{j,k} \subset \mathcal{X}$, whose left and right siblings respectively depict disjoint subsets $\mathcal{C}_{j+1,2k}$ and $\mathcal{C}_{j+1,2k+1}$ such that $\mathcal{C}_{j,k} = \mathcal{C}_{j+1,2k} \cup \mathcal{C}_{j+1,2k+1}$. At the root, one starts with a constant scoring function $s_1(x) = \mathbb{I}\{x \in \mathcal{C}_{0,0}\} \equiv 1$ and after $m = 2^j + k$ iterations, $0 \leq k < 2^j$, the current scoring function is $s_m(x) = \sum_{l=0}^{2k-1}(m-l) \cdot \mathbb{I}\{x \in \mathcal{C}_{j+1,l}\} + \sum_{l=k}^{2^j-1}(m-k-l) \cdot \mathbb{I}\{x \in \mathcal{C}_{j,l}\}$ and the cell $\mathcal{C}_{j,k}$ is split in order to form an updated version of the scoring function, $s_{m+1}(x) = \sum_{l=0}^{2k}(m-l) \cdot \mathbb{I}\{x \in \mathcal{C}_{j+1,l}\} + \sum_{l=k+1}^{2^j-1}(m-k-l) \cdot \mathbb{I}\{x \in \mathcal{C}_{j,l}\}$ namely, with maximum (empirical) AUC. Therefore, it happens that this problem boils down to solve a cost-sensitive binary classification problem on the set $\mathcal{C}_{j,k}$, see subsection 3.3 in Clémençon et al. (2011) for further details. Indeed, one may write the AUC increment as

$$\mathrm{AUC}_{1,2}(s_{m+1}) - \mathrm{AUC}_{1,2}(s_m) = \frac{1}{2} F_1(\mathcal{C}_{j,k}) F_2(\mathcal{C}_{j,k})(1 - \Lambda_{1,2}(\mathcal{C}_{j+1,2k} \mid \mathcal{C}_{j,k})),$$

where

$$\Lambda_{1,2}(\mathcal{C}_{j+1,2k} \mid \mathcal{C}_{j,k}) \stackrel{def}{=} F_2(\mathcal{C}_{j,k} \setminus \mathcal{C}_{j+1,2k})/F_2(\mathcal{C}_{j,k}) + F_1(\mathcal{C}_{j+1,2k})/F_1(\mathcal{C}_{j,k}).$$

Setting $p = F_2(\mathcal{C}_{j,k})/(F_1(\mathcal{C}_{j,k}) + F_2(\mathcal{C}_{j,k}))$, the crucial point of the TreeRank approach is that the quantity $2p(1-p)\Lambda_{1,2}(\mathcal{C}_{j+1,2k} \mid \mathcal{C}_{j,k})$ can be seen as the cost-sensitive error[1] of a classifier on $\mathcal{C}_{j,k}$ predicting label 2 on $\mathcal{C}_{j+1,2k}$ and label 1 on $\mathcal{C}_{j,k} \setminus \mathcal{C}_{j+1,2k}$ with cost p (respectively, $1-p$) assigned to the error consisting in predicting label 2 given $Y = 1$ (resp., label 1 given $Y = 2$), balancing thus the two types of error. Hence, at each iteration of the ranking tree growing stage, the TreeRank algorithm calls a *cost-sensitive* binary classification algorithm, termed LeafRank, in order to solve a statistical version of the problem above (replacing the theoretical probabilities involved by their empirical counterparts) and split $\mathcal{C}_{j,k}$ into $\mathcal{C}_{j+1,2k}$ and $\mathcal{C}_{j+1,2k+1}$. As described at length in Clémençon et al. (2011), one may use cost-sensitive versions of celebrated binary classification algorithms such as CART or SVM for instance as LeafRank procedure, the performance depending on their ability to capture the geometry of the level sets of the likelihood ratio $dF_2/dF_1(x)$. The procedure is depicted in Fig. 1. In general, the growing stage is followed by a pruning procedure, where children of a same parent node are recursively merged in order to produce a ranking subtree that maximizes an estimate of the AUC criterion, based on cross-validation usually (*cf* section 4 in Clémençon et al. (2011)). Under adequate assumptions, consistency results and rate bounds for the TreeRank approach (in the sup norm sense and for the AUC deficit both at the same time) are established in Clémençon and Vayatis (2009) and Clémençon et al. (2011), an extensive experimental study can be found in Clémençon et al. (2012).

2.4. *Multipartite Ranking Algorithms*

In contrast to the bipartite situation (see Clémençon and Vayatis (2010), Clémençon and Vayatis (2009)), no algorithm optimizing the ROC surface directly and producing a scoring function $\widehat{s}_n$ for which $d_\infty(\widehat{s}_n, s^*) \to 0$ in probability has been documented in the

[1] Let (X', Y') be a random pair, where Y' takes binary values, in $\{1,\ 2\}$ say, and X' models some information valued in a space $\mathcal{X}'$, hopefully useful to predict the label Y'. A classifier is any measurable mapping $g : \mathcal{X}' \to \{1,\ 2\}$. Let $p' = \mathbb{P}\{Y = 2\}$. Given a cost $\omega \in [0, 1]$, the cost-sensitive error of g is $L_\omega(g) \stackrel{def}{=} 2p'(1-\omega)\mathbb{P}\{g(X') = 1 \mid Y = 2\} + 2(1-p')\omega\mathbb{P}\{g(X') = 2 \mid Y = 1\}$. The quantity $L_{1/2}(g)$ is generally referred to as the error of classifier g.

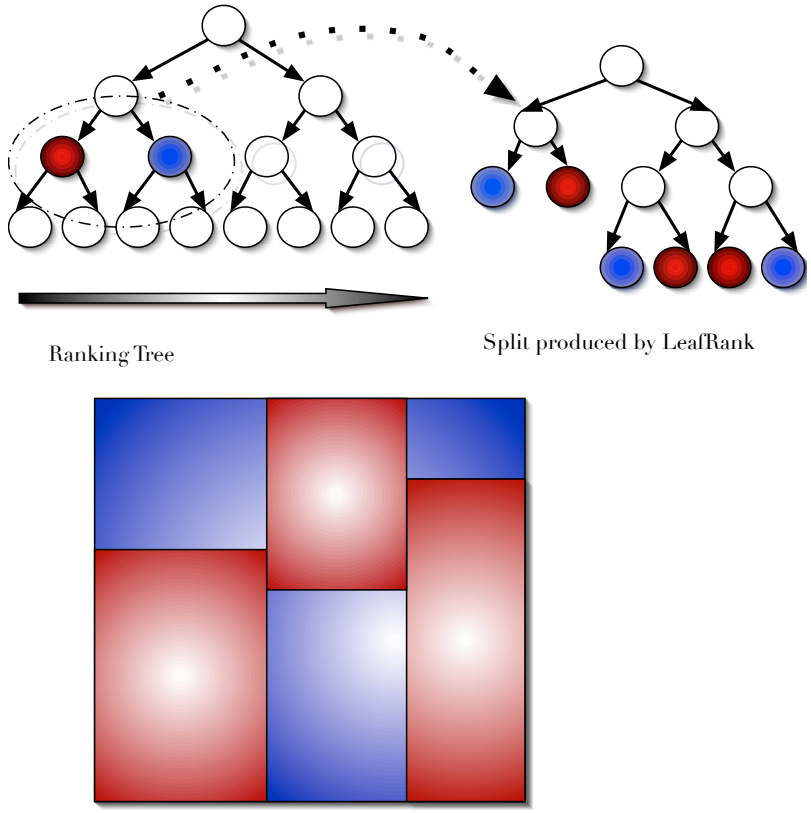


Figure 1. Ranking Tree with LeafRank splits built through a cost-sensitive version of the CART algorithm with a local cost depending on the rate of "positive" instances within the node to be split

literature. Beyond theoretical results guaranteeing the validity of empirical maximization of the VUS criterion (see Rajaram and Agarwal (2005)), most methods proposed rely on the optimization of an alternative (pairwise) criterion (Freund et al. (2003) and Pahikkala et al. (2007)), or on the decomposition of the original multipartite problem into bipartite subproblems combined with a final aggregation/consensus stage (Hüllermeier et al. (2008) and Clémençon et al. (2013)) or still on plug-in approaches based on ordinal regression (Waegeman et al. (2008)). In addition, it is far from straightforward to extend the TreeRank algorithm recalled above because, when $K \geq 3$, as a straightforward computation based on Eq. (1) may show, the splitting step cannot be interpreted as a learning problem which can be solved by means of off-the-shelf techniques, unlike the bipartite case. Indeed, taking $s(x) = \mathbb{I}\{x \in \mathcal{C}\}$ for some measurable set $\mathcal{C} \subset \mathcal{X}$, we have

$$\mathrm{VUS(s)} = \mathrm{F}_3(\mathcal{C})(1 - \mathrm{F}_1(\mathcal{C}))/2 + (1 - \mathrm{F}_1(\mathcal{C}))(1 - \mathrm{F}_2(\mathcal{C}))(1 - \mathrm{F}_3(\mathcal{C}))/6 + F_1(\mathcal{C})F_2(\mathcal{C})F_3(\mathcal{C})/6. \quad (2)$$

It is the goal of this paper to propose an alternative, letting splitting rule candidates, corresponding to solutions of different bipartite subproblems, compete for VUS maximization in a tournament.

3. The TreeRank Tournament Algorithm

We now describe the algorithm we propose to solve the multipartite ranking problem. We place ourselves in the tripartite case for notational simplicity, but extension to the general multipartite setting is straightforward, *cf* Appendix A. The algorithm is implemented from a training dataset $\mathcal{D}_n = \{(X_1,\ Y_1),\ \ldots,\ (X_n,\ Y_n)\}$ and recursively calls a *cost-sensitive* binary classification algorithm $\mathcal{L}$ (*e.g.* SVM, CART, Random Forest, k-NN), referred to as LeafRank. When ran on a set $\mathcal{C} \subset \mathcal{X}$, we denote by $\mathcal{L}(\mathcal{C})$ the collection of subsets of $\mathcal{C}$ over which algorithm $\mathcal{L}$ performs optimization. For $1 \leq k \leq 3$, set $n_k = \sum_{i=1}^n \mathbb{I}\{Y_i = k\}$ and define, for any measurable set $\mathcal{C} \subset \mathcal{X}$, $\widehat{F}_k(\mathcal{C}) = (1/n_k)\sum_{i=1}^n \mathbb{I}\{X_i \in \mathcal{C},\ Y_i = k\}$, and, for any measurable subset $\mathcal{C}' \subset \mathcal{C}$ with $1 \leq k < l \leq 3$,

$$\widehat{\Lambda}_{k,l}(\mathcal{C}' \mid \mathcal{C}) = \widehat{F}_l(\mathcal{C}')/\widehat{F}_l(\mathcal{C}) + \widehat{F}_k(\mathcal{C} \setminus \mathcal{C}')/\widehat{F}_k(\mathcal{C}).$$

As already pointed out in subsection 2.3, the quantity above can be seen as proportional to the empirical *cost-sensitive* error of a binary classifier on the restricted input space $\mathcal{C}$ which predicts label l on $\mathcal{C}'$ and label k on $\mathcal{C} \setminus \mathcal{C}'$ with cost $\widehat{F}_l(\mathcal{C})/(\widehat{F}_k(\mathcal{C}) + \widehat{F}_l(\mathcal{C}))$ (respectively, $\widehat{F}_k(\mathcal{C})/(\widehat{F}_k(\mathcal{C}) + \widehat{F}_l(\mathcal{C}))$) assigned to the error consisting in predicting label l while the true label is k (resp., label k, while the true label is l), based on the data of the original sample $\mathcal{D}_n$ lying in the set $\mathcal{C}$ with label k or l. We also introduce the quantity:

$$\begin{aligned}\widehat{\mathrm{VUS}}_{\mathcal{C}}(\mathcal{C}') = \widehat{F}_3(\mathcal{C}')(\widehat{F}_1(\mathcal{C}) - \widehat{F}_1(\mathcal{C}'))/2 + \widehat{F}_1(\mathcal{C}')\widehat{F}_2(\mathcal{C}')\widehat{F}_3(\mathcal{C}')/6 \\ + (\widehat{F}_1(\mathcal{C}) - \widehat{F}_1(\mathcal{C}'))(\widehat{F}_2(\mathcal{C}) - \widehat{F}_2(\mathcal{C}'))(\widehat{F}_3(\mathcal{C}) - \widehat{F}_3(\mathcal{C}'))/6,\end{aligned}$$

which corresponds to the empirical VUS increase resulting from splitting the cell $\mathcal{C}$ into left and right siblings $\mathcal{C}'$ and $\mathcal{C} \setminus \mathcal{C}'$, *cf* Eq. (2).

A straightforward variant of the TreeRank Tournament algorithm could consist in running additionally the LeafRank algorithm for local cost-sensitive classification problems related to the pair of labels $(1, 3)$ and thus enlarging the set of competitors ("extended tournament"). This would however increase the amount of computations performed. In addition, just like for TreeRank algorithm in the bipartite context (see Clémençon et al. (2011)) and for most other recursive partitioning methods, the ranking tree growing procedure can be followed by a pruning stage, where children of a same parent node can be merged recursively in order to maximize a (cross-validation based) estimate of the VUS criterion. Bootstrap aggregating techniques relying on concepts pertaining to the *ranking consensus* theory and randomization could also be considered to design committee based multipartite ranking rules, possibly improving over single ranking trees, as in Clémençon et al. (2013). Model selection analysis and aggregation are however beyond the scope of the present article and will be dealt with in a future work.

We also highlight the advantage of the TreeRank Tournament algorithm regarding missing data: they can be handled in a straightforward fashion by assigning to a partially observed instance x the empirical mean of each unobserved component within the cell where it currently lies in the training stage or for prediction. Another advantage of decision trees lies in their interpretability. Indeed, a ranking tree may be easily visualized in two dimensions, see Fig. 1 and the related scoring function may be described through a chain of simple rules. When designing medical diagnosis supporting tools or credit-scoring

TreeRank Tournament

(1) (Input.) Training sample $\mathcal{D}_n$, LeafRank algorithm $\mathcal{L}$, ranking tree depth J.

(2) (Initialization.) Set $\mathcal{C}_{0,0} = \mathcal{X}$ and $s_0(x) \equiv 1$.

(3) (Iterations.) For $m = 1, \ldots, 2^J$, define $j = \langle \log m / \log 2 \rangle$ and $l = m - 2^j$, and then

a) (LeafRank runs.) For all $k \in \{1, 2\}$, run algorithm $\mathcal{L}$ in order to output

$$\widetilde{\mathcal{C}}^{(k)} = \operatorname*{arg\,max}_{\mathcal{C} \in \mathcal{L}(\mathcal{C}_{j,l})} \widehat{\Lambda}_{k,k+1}(\mathcal{C} \mid \mathcal{C}_{j,l}).$$

b) (Tournament.) Compute

$$\mathcal{C}_{j+1,2l} = \operatorname*{arg\,max}_{\widetilde{\mathcal{C}}^{(k)},\, k=1,\,2} \widehat{\mathrm{VUS}}_{\mathcal{C}_{d,l}}(\widetilde{\mathcal{C}}^{(k)}),$$

and set $\mathcal{C}_{j+1,2l+1} = \mathcal{C}_{j,l} \setminus \mathcal{C}_{j+1,2l}$.

(4) (Output.) Compute the piecewise constant scoring function :

$$s_{2^J}(x) = \sum_{l=0}^{2^J-1} (2^J - l) \cdot \mathbb{I}\{x \in \mathcal{C}_{J,l}\}.$$

rules in banking for instance, it is essential to interpret the score s(x) and determine which attributes contribute the most to its variation. Possible monitoring tools (*e.g.* variable importance, partial dependence plots) can be immediately deduced from those discussed in section 5 of Clémençon et al. (2011) in the bipartite case, replacing the AUC criterion by the VUS criterion.

Before providing theoretical guarantees for the TreeRank Tournament algorithm of the form of rate bounds for $d_i(s_{2^J}, s^*)$ with $i \in \{1, \infty\}$, we shall now analyze an adaptive piecewise linear approximation scheme to recover the optimal ROC surface with a controlled error rate, which is somehow mimicked by the learning algorithm above.

4. Adaptive Piecewise Planar Approximation of ROC*

This section is dedicated to the analysis of an adaptive approximation scheme of the optimal ROC surface, which outputs a piecewise planar approximate of ROC* that is itself the ROC surface of a piecewise constant scoring function. In order to describe it at length, further notations are required.

4.1. *Further Notations and Preliminaries*

Let $\mathcal{P} = (\mathcal{C}_j)_{1 \leq j \leq N}$ be an *ordered partition* of the input space $\mathcal{X}$ counting $N \geq 1$ cells. The adjective *ordered* means here that, for any $1 \leq i \leq j \leq N$, instances lying in $\mathcal{C}_i$ are expected to have higher labels than those in $\mathcal{C}_j$, in a way that $\mathcal{P}$ is related to the scoring

function

$$S_{\mathcal{P}}(x) = \sum_{i=1}^{N} (N - i + 1) \cdot \mathbb{I}\{x \in \mathcal{C}_i\}.$$

We point out that in the tripartite case, $S_{\mathcal{P}}$'s ROC surface is piecewise planar with N^2 pieces. More precisely, it is the polytope that connects the points

$$\left(1 - \sum_{l=1}^{i} F_1(\mathcal{C}_l),\ \sum_{l=j+1}^{i} F_2(\mathcal{C}_l),\ \sum_{l=1}^{j} F_3(\mathcal{C}_l) \right),$$

where $0 \leq j \leq i \leq N$, with the convention that empty summation equals zero. In order to provide a closed analytical form for the latter, set $\alpha_j = 1 - F_1(\cup_{l=1}^{j} \mathcal{C}_l)$ and $\gamma_j = F_3(\cup_{l=1}^{j} \mathcal{C}_l)$ for $j = 1, \ldots, N$ and $1 - \alpha_0 = \alpha_{N+1} = \gamma_0 = 1 - \gamma_{N+1} = 0$ by convention. Set also

$$\phi(\alpha, \alpha', \alpha'') = \frac{\alpha - \alpha'}{\alpha'' - \alpha'} \mathbb{I}\{\alpha \in [\alpha'; \alpha'']\}$$

for all $\alpha' \leq \alpha \leq \alpha''$ and consider the *hat functions* defined by

$$\phi_i(\alpha) = \phi(\alpha, \alpha_{i-1}, \alpha_i) - \phi(\alpha, \alpha_i, \alpha_{i+1}),$$
$$\varphi_j(\gamma) = \phi(\gamma, \gamma_{j-1}, \gamma_j) - \phi(\gamma, \gamma_j, \gamma_{j+1}),$$

as well as the tensorial products $\Phi_{i,j}(\alpha, \gamma) = \phi_i(\alpha)\varphi_j(\gamma)$ for $1 \leq i,\ j \leq N$, which are the basis functions used in the *Finite Element Method* to approximate real valued functions defined on $[0, 1]^2$. Equipped with these notations, the ROC surface of $S_{\mathcal{P}}$ can be written as

$$\mathrm{ROC}_{S_{\mathcal{P}}}(\alpha, \gamma) = \sum_{1 \leq j \leq i \leq N} F_2 \left(\bigcup_{l=1+j}^{i} \mathcal{C}_l \right) \Phi_{i,j}(\alpha, \gamma). \tag{3}$$

4.2. *An Implicit Tree-Structured Recursive Interpolation Scheme*

Here, we describe a recursive approximation scheme to build a piecewise constant scoring function $S^*_{\mathcal{P}^*}$ whose ROC surface can be viewed as a piecewise planar interpolant of ROC^*, corresponding to a mesh grid adaptively chosen. As shall be seen below, the related oriented partition $\mathcal{P}^*$ can be represented by means of a left-to-right oriented binary tree structure $\{\mathcal{C}^*_{j,l} :\ j \leq J,\ l = 0, \ldots, 2^j - 1\}$ and its cells coincide with certain bilevel sets of the regression function $\eta(x)$. In addition, as shall be seen below, the distance (in sup-norm) between $\mathrm{ROC}_{S^*_{\mathcal{P}^*}}$ and ROC^* can be bounded as a function of the number of iterations (*i.e.* of the number of cells of $\mathcal{P}^*$) under the following assumptions.

Assumption 4.1: The class distributions F_1, F_2 and F_3 are equivalent and the likelihood ratios $\Phi_{2,1}, \Phi_{3,1}, \Phi_{3,2}$ are bounded.

Assumption 4.2: The distribution of $\eta(X)$ is absolutely continuous with respect to Lebesgue measure. Let $F_k^*(x)$ and $F_k^*(dx) = f_k^*(x)dx$ be the conditional cdf and df of $\eta(X)$ given $Y = k$, with $1 \leq k \leq 3$.

In particular, these hypotheses guarantee that the optimal ROC surface exhibits a minimum amount of smoothness, as stated in the proposition below.

Proposition 4.3: *Under Condition 2.1, Assumptions 4.1 and 4.2, the mapping* $(\alpha, \gamma) \in [0,1]^2 \mapsto \mathrm{ROC}^*(\alpha, \gamma)$ *is differentiable. On the set* $\mathcal{I}^* = \{(\alpha,\ \gamma) \in [0,1]^2 : F_2^* \circ F_3^{*-1}(1-\gamma) \geq F_2^* \circ F_1^{*-1}(\alpha)\}$*, the first partial derivatives of* ROC^* *are given by:*

$$\frac{\partial}{\partial \alpha}\mathrm{ROC}^*(\alpha,\gamma) = -\frac{f_2^*}{f_1^*}(F_1^{*-1}(\alpha)),\ \ \frac{\partial}{\partial \gamma}\mathrm{ROC}^*(\alpha,\gamma) = -\frac{f_2^*}{f_3^*}(F_1^{*-1}(1-\gamma)).$$

They are equal to zero on the complementary set $[0,1]^2 \setminus \mathcal{I}^*$.

The subsequent analysis actually actually requires that a slightly stronger smoothness assumption holds true.

Assumption 4.4: The mapping ROC^* is twice differentiable with bounded second derivatives given by: $\forall (\alpha, \gamma) \in \mathcal{I}^*$,

$$\frac{\partial^2}{\partial \alpha^2}\mathrm{ROC}^*(\alpha,\gamma) = -\frac{f_2'^* f_1^* - f_2^* f_1'^*}{f_1^{*3}}(F_1^{*-1}(\alpha)),$$
$$\frac{\partial^2}{\partial \gamma^2}\mathrm{ROC}^*(\alpha,\gamma) = \frac{f_2'^* f_3^* - f_2^* f_3'^*}{f_3^{*3}}(F_3^{*-1}(1-\gamma)).$$

We now describe at length the approximation scheme.
Initialization. We set $\mathcal{C}_{0,0}^* = \mathcal{X}$, $s_1^*(x) \equiv 1$ and $1 = \alpha_{0,0}^* = 1 - \alpha_{0,1}^* = 1 - \gamma_{0,0}^* = \gamma_{0,1}^* = 1 - \beta_{0,0}^* = \beta_{0,1}^*$. Observe that $F_1(\mathcal{C}_{0,0}^*) = \alpha_{0,0}^* - \alpha_{0,1}^*$, $F_2(\mathcal{C}_{0,0}^*) = \beta_{0,1}^* - \beta_{0,0}^*$ and $F_3(\mathcal{C}_{0,0}^*) = \gamma_{0,1}^* - \gamma_{0,0}^*$. In the $\alpha\gamma\beta$ system of coordinates, the initial approximant of the surface ROC^* is the planar piece connecting $(1,0,0) = (\alpha_{0,0}^*, \gamma_{0,0}^*, \beta_{0,0}^*)$, $(0,1,0) = (\alpha_{0,1}^*, \gamma_{0,1}^*, \beta_{0,0}^*)$ and $(0,0,1) = (\alpha_{0,1}^*, \gamma_{0,0}^*, \beta_{0,1}^*)$. It is the surface $\{(\alpha,\ \gamma, \widetilde{\mathrm{ROC}}_1^*(\alpha,\gamma)) :\ (\alpha,\gamma) \in [0,1]^2\}$ with $\widetilde{\mathrm{ROC}}_1^*(\alpha,\gamma) = 1 - \alpha - \gamma$.

Iterations. For $j = 0,\ \ldots,\ J-1$ and for $k = 0,\ \ldots,\ 2^j - 1$:

• **Updates.** Set $\alpha_{j+1,2k}^* = \alpha_{j,k}^*$, $\alpha_{j+1,2k+2}^* = \alpha_{j,k+1}^*$, $\gamma_{j+1,2k}^* = \gamma_{j,k}^*$ and $\gamma_{j+1,2k+2}^* = \gamma_{j,k+1}^*$, $\beta_{j+1,2k}^* = \beta_{j,k}^*$ and $\beta_{j+1,2k+2}^* = \beta_{j,k+1}^*$.

• **Breakpoint candidates.** Considering the curve formed by the intersection between the current approximant of ROC^* and the facet "$\gamma = 0$", define the point of coordinate

$$\alpha_{j+1,2k+1}^{(1)} = \mathrm{ROC}_{1,2}'^{*-1}\left(\frac{\beta_{j,k+1}^* - \beta_{j,k}^*}{\alpha_{j,k}^* - \alpha_{j,k+1}^*}\right)$$

on the α axis. This corresponds to the largest increase of the area under the curve when adding a breakpoint between $\alpha_{j,k}^*$ and $\alpha_{j,k+1}^*$, see Proposition 11 in

Clémençon and Vayatis (2009). Incidentally, the resulting broken line is also optimal in the sup norm sense. Observe also that $\alpha^{(1)}_{j+1,2k+1} = \alpha^*_{j+1,2k} - F_1(\mathcal{C}^{(1)}_{j+1,2k})$, where $\mathcal{C}^{(1)}_{j+1,2k} = \arg\max_{\mathcal{C}\subset\mathcal{C}_{j,k}} \Lambda_{1,2}(\mathcal{C} \mid \mathcal{C}_{j,k})$. In addition, we have $\mathcal{C}^{(1)}_{j+1,2k} = \{x \in \mathcal{X} : F^{-1}_{\Phi_{1,2},1}(\alpha_{j+1,2k+1}) < \Phi_{1,2}(x) \leq F^{-1}_{\Phi_{1,2},1}(\alpha_{j+1,2k})\}$, where $F^{-1}_{\Phi_{1,2},1}(\alpha)$ denotes the quantile of order α of $\Phi_{1,2}(X)$'s conditional distribution given $Y = 1$. We also set $\beta^{(1)}_{j+1,2k+1} = \beta^*_{j+1,2k} + F_2(\mathcal{C}^{(1)}_{j+1,2k}) = \mathrm{ROC}^*_{1,2}(1 - \alpha^{(1)}_{j+1,2k+1})$ and $\gamma^{(1)}_{j+1,2k+1} = \gamma^*_{j+1,2k} + F_3(\mathcal{C}^{(1)}_{j+1,2k}) = \mathrm{ROC}^*_{1,3}(1 - \alpha^{(1)}_{j+1,2k+1})$.

In the same fashion, considering the curve formed by the intersection between the current approximate of ROC^* and the facet "$\alpha = 0$", define the point of coordinate

$$\gamma^{(2)}_{j+1,2k+1} = \mathrm{ROC}'^{*-1}_{2,3}\left(\frac{\gamma^*_{j,k+1} - \gamma^*_{j,k}}{\beta^*_{j,k+1} - \beta^*_{j,k}}\right)$$

on the γ axis. This corresponds to the largest increase of the area under the curve when adding a breakpoint between $\gamma^*_{j,k}$ and $\gamma^*_{j,k+1}$. We have $\gamma^{(2)}_{j+1,2k+1} = \gamma^*_{j+1,2k} + F_3(\mathcal{C}^{(2)}_{j+1,2k})$, where $\mathcal{C}^{(2)}_{j+1,2k} = \arg\max_{\mathcal{C}\subset\mathcal{C}_{j,k}} \Lambda_{2,3}(\mathcal{C} \mid \mathcal{C}_{j,k})$. In addition, we have $\mathcal{C}^{(2)}_{j+1,2k} = \{x \in \mathcal{X} : F^{-1}_{\Phi_{2,3},3}(\gamma_{j+1,2k+1}) < \Phi_{2,3}(x) \leq F^{-1}_{\Phi_{2,3},1}(\gamma_{j+1,2k})\}$, where $F^{-1}_{\Phi_{2,3},3}(\gamma)$ denotes the quantile of order γ of $\Phi_{2,3}(X)$'s conditional distribution given $Y = 3$. We also set $\alpha^{(2)}_{j+1,2k+1} = \alpha^*_{j+1,2k} - F_1(\mathcal{C}^{(2)}_{j+1,2k}) = 1 - \mathrm{ROC}^*_{3,1}(\gamma^{(2)}_{j+1,2k+1})$ and $\beta^{(2)}_{j+1,2k+1} = \beta^*_{j+1,2k} + F_2(\mathcal{C}^{(2)}_{j+1,2k}) = \mathrm{ROC}^*_{3,2}(\gamma^{(2)}_{j+1,2k+1})$.

• **Tournament.** For $l \in \{1,\ 2\}$, compute the quantity

$$\begin{aligned}
\mathrm{VUS}_{\mathcal{C}^*_{\mathrm{j,k}}}(\mathcal{C}^{(\mathrm{l})}_{\mathrm{j+1,2k}}) &= \mathrm{F}_3(\mathcal{C}^{(\mathrm{l})}_{\mathrm{j+1,2k}})(\mathrm{F}_1(\mathcal{C}^*_{\mathrm{j,k}}) - \mathrm{F}_1(\mathcal{C}^{(\mathrm{l})}_{\mathrm{j+1,2k}}))/2 \\
&\quad + F_1(\mathcal{C}^{(l)}_{j+1,2k})F_2(\mathcal{C}^{(l)}_{j+1,2k})F_3(\mathcal{C}^{(l)}_{j+1,2k})/6 \\
&\quad + (F_1(\mathcal{C}^*_{j,k}) - F_1(\mathcal{C}^{(l)}_{j+1,2k}))(F_2(\mathcal{C}^*_{j,k}) - F_2(\mathcal{C}^{(l)}_{j+1,2k}))(F_3(\mathcal{C}^*_{j,k}) - F_3(\mathcal{C}^{(l)}_{j+1,2k}))/6 \\
&= (\gamma^{(l)}_{j+1,2k+1} - \gamma^*_{j+1,2k})(\alpha^{(l)}_{j+1,2k+1} - \alpha^*_{j+1,2k+2})/2 \\
&\quad + (\alpha^*_{j+1,2k} - \alpha^{(l)}_{j+1,2k+1})(\beta^{(l)}_{j+1,2k+1} - \beta^*_{j+1,2k})(\gamma^{(l)}_{j+1,2k+1} - \gamma^*_{j+1,2k})/6 \\
&\quad + (-\alpha^*_{j+1,2k+2} + \alpha^{(l)}_{j+1,2k+1})(-\beta^{(l)}_{j+1,2k+1} + \beta^*_{j+1,2k+2})(-\gamma^{(l)}_{j+1,2k+1} + \gamma^*_{j+1,2k+2})/6.
\end{aligned}$$

Then, determine

$$l^* = \underset{l=1,\ 2}{\arg\max}\, \mathrm{VUS}_{\mathcal{C}^*_{\mathrm{j,k}}}(\mathcal{C}^{(\mathrm{l})}_{\mathrm{j+1,2k}})$$

and set $\mathcal{C}^*_{j+1,2k} = \mathcal{C}^{(l^*)}_{j+1,2k}$ and $\mathcal{C}^*_{j+1,2k+1} = \mathcal{C}^*{}_{j,k} \setminus \mathcal{C}^{(l^*)}_{j+1,2k}$. Fig. 2 below depicts this step of the approximation scheme. In addition, define $\alpha^*_{j+1,2k+1} = \alpha^*_{j+1,2k} - F_1(\mathcal{C}^*_{j+1,2k})$, $\beta^*_{j+1,2k+1} = \beta^*_{j+1,2k} + F_2(\mathcal{C}^*_{j+1,2k})$ and $\gamma^*_{j+1,2k+1} = \gamma^*_{j+1,2k} + F_3(\mathcal{C}^*_{j+1,2k})$.

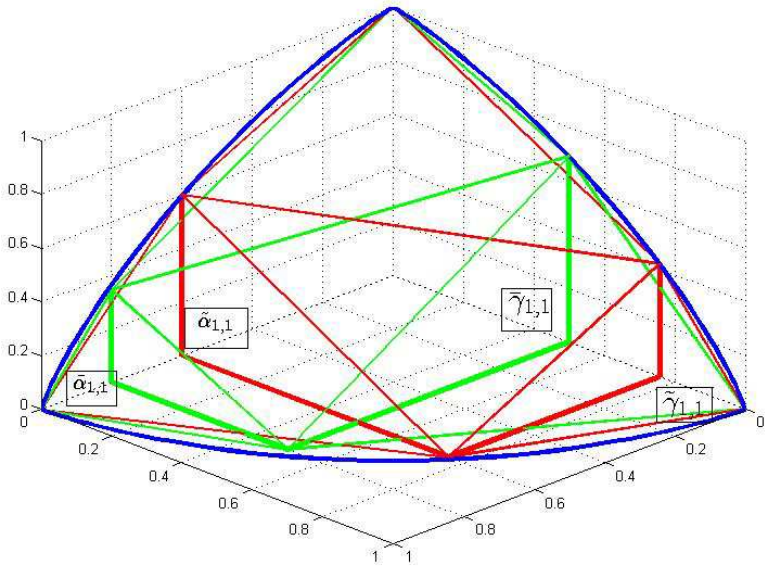


Figure 2. Tournament selection of the "best" breakpoint

Output. Compute the approximate given by: $\forall(\alpha,\gamma)\in[0,1]^2$,

$$\widetilde{\mathrm{ROC}}_J^*(\alpha,\gamma)=\sum_{1\leq l\leq i\leq 2^J-1}\left(\beta^*_{J,i}-\beta^*_{J,l}\right)\Phi^*_{i,l}(\alpha,\gamma),$$

where, for $1\leq i,\ l\leq 2^J-1$, we have set $\Phi_{i,l}(\alpha,\gamma)=\phi_i^*(\alpha)\varphi_l^*(\gamma)$ with $\phi_i^*(\alpha)=\phi(\alpha,\alpha^*_{J,i-1},\alpha^*_{J,i})-\phi(\alpha,\alpha^*_{J,i},\alpha^*_{J,i+1})$ and $\varphi_l(\gamma)=\phi(\gamma,\gamma^*_{J,l-1},\gamma^*_{J,l})-\phi(\gamma,\gamma^*_{J,l},\gamma^*_{J,l+1})$. Observe that it is the ROC surface of the scoring function:

$$s^*_{2^J}(x)=\sum_{l=0}^{2^J-1}(2^J-l)\cdot\mathbb{I}\{x\in\mathcal{C}^*_{J,l}\}.$$

Indeed, we have: $\widetilde{\mathrm{ROC}}_J^*(\alpha,\gamma)=\mathrm{ROC}_{s^*_J}(\alpha,\gamma)$ for all $(\alpha,\gamma)\in[0,1]^2$.

It is noteworthy that the interpolant of the optimal ROC surface produced by the algorithm above is itself a (concave) ROC surface. Obviously, this is not the case in general, *cf* Eq. (3) above. This strikingly differs from the bipartite case, where any interpolant of the optimal ROC curve is the ROC curve of a piecewise constant scoring function, constant on certain bilevel sets of the likelihood ratio related to the class distributions, see subsection 3.1 in Clémençon and Vayatis (2010).

The following result provides guarantees for the approximation scheme described above.

Proposition 4.5: *Under Condition 2.1, Assumptions 4.1, 4.2 and 4.4, there exists a constant $C<+\infty$ such that:*

$$\forall J\geq 1,\ \ d_\infty(s^*,s^*_{2^J})\leq C\times 2^{-2J}.$$

Now, the TreeRank Tournament algorithm can be clearly viewed as a statistical version of the interpolation scheme above. It will mimic it well, provided that each tournament yields a splitting rule closed to that based on the true VUS increment. This is the key to establish the rate bounds displayed in the next section.

5. Main results

It is the goal of this section to display results of statistical and empirical nature, so that the TreeRank Tournament algorithm can be grounded in a strong validity framework. Beyond an analysis of its generalization ability, numerical experiments have been carried out in order to compare the performance of the method proposed to that of alternative techniques documented in the literature.

5.1. *Learning rate bounds*

The following noise assumption, used in Clémençon et al. (2013) and generalizing that introduced in Clémençon et al. (2008) in the bipartite setup, shall be involved in the analysis.

Assumption 5.1: For $k \in \{1, 2\}$, the (pairwise) posterior probability given by $\eta_{k+1}(X)/(\eta_k(X) + \eta_{k+1}(X))$ is a continuous random variable and there exist $c < \infty$ and $a \in (0, 1)$ such that

$$\forall x \in \mathcal{X}, \ \mathbb{E}\left[\left|\frac{\eta_{k+1}(X)}{\eta_{k+1}(X) + \eta_k(X)} - \frac{\eta_{k+1}(x)}{\eta_{k+1}(x) + \eta_k(x)}\right|^{-a}\right] \leq c \ . \tag{4}$$

As revealed by the theorem below, equipped with this additional hypothesis, one may connect the performance of the splitting rule winner of the empirical tournament to that of the winner of the tournament based on the true VUS increment. The result is then established by following line by line the argument of Theorem 15 in Clémençon and Vayatis (2009), see the sketch of proof given in the Appendix section.

Theorem 5.2: *Assume that Condition 2.1, Assumptions 4.1, 4.2, 4.4 and 5.1 hold. Suppose that the class $\mathcal{L}(\mathcal{X})$ of subsets candidates is of finite* VC *dimension V, contains all level sets $\{x \in \mathcal{X} : \eta(x) \geq t\}$, $t \in \mathbb{R}$, of the regression function (or of optimal scoring functions equivalently) and that $\mathcal{L}(\mathcal{X}) \cap \mathcal{C} = \mathcal{L}(\mathcal{C})$ for all $\mathcal{C} \in \mathcal{L}(\mathcal{X})$. Then, there exists a constant c_0 and universal constants c_1 and c_2 such that, for all $\delta > 0$, with probability at least $1 - \delta$, we have: for all $J \geq 1$ and $n \geq 1$,*

$$d_1(s_{2^J}, s^*_{2^J}) \leq c_0^J \left\{ \left(c_1^2 V/n\right)^{\frac{a^J}{2(1+a)^J}} + \left(c_2^2 \log(1/\delta)/n\right)^{\frac{a^J}{2(1+a)^J}} \right\}.$$

Combined with Proposition 4.5, the result stated above provides rate bounds in the ROC space. Naturally, because of the hierarchical structure of the oriented partition produced by the TreeRank Tournament algorithm, slow rate bounds were expected. We point out however that the bounds exhibited hold true under very general assumptions and correspond to confidence regions in sup norm (analogous results in terms of VUS immediately follow).

5.2. *Experimental results*

We now investigate the numerical performance of the TreeRank Tournament algorithm (referred to as "TRT" in the tables below), on toy and real datasets.

Here, the LeafRank procedure is a locally weighted version of the algorithm CART. Routines of the R package available at http://treerank.sourceforge.net, are used to implement TreeRank Tournament (with "default" parameters minsplit= 1 , maxdepth= 4). For comparison purpose, we also ran ranking algorithms, standing as natural competitors: RankBoost (when aggregating 30 stumps, see Rudin et al. (2005), called "RBpw") and SVMRank (with linear and Gaussian kernels with respective parameters $C = 20$ and $(C, \gamma) = (0.01)$, see Herbrich et al. (2000), called respectively "SVMl" and "SVMg"), using the SVM-light implementation available at http://svmlight.joachims.org/. We have also used the RankRLS method (http://www.tucs.fi/RLScore, see Pahikkala et al. (2007), called respectively "RLSl" and "RLSg") that implements a regularized least square algorithm with linear kernel ("$bias = 1$") and with Gaussian kernel ($\gamma = 0.01$), selection of the intercept on a grid being performed through a leave-one-out procedure.

Mixtures of Gaussian distributions. Consider Z a q-dimensional random vector from a Gaussian distribution drawn $\mathcal{N}(\mu, \Gamma)$, and a Borelian set $C \subset \mathbb{R}^q$. We denote by $\mathcal{N}_C(\mu, \Gamma)$ the conditional distribution of Z given $Z \in C$. Equipped with this notation, we can write the class distributions used in this example as:

$$\phi_1(x) = \mathcal{N}_{[0,1]^2}\left(\begin{pmatrix}0\\0\end{pmatrix}, \begin{pmatrix}1/4 & 0\\0 & 1/4\end{pmatrix}\right)$$
$$\phi_2(x) = \mathcal{N}_{[0,1]^2}\left(\begin{pmatrix}1/2\\1/2\end{pmatrix}, \begin{pmatrix}1/4 & 0\\0 & 1/4\end{pmatrix}\right)$$
$$\phi_3(x) = \mathcal{N}_{[0,1]^2}\left(\begin{pmatrix}1\\1\end{pmatrix}, \begin{pmatrix}1/4 & 0\\0 & 1/4\end{pmatrix}\right)$$

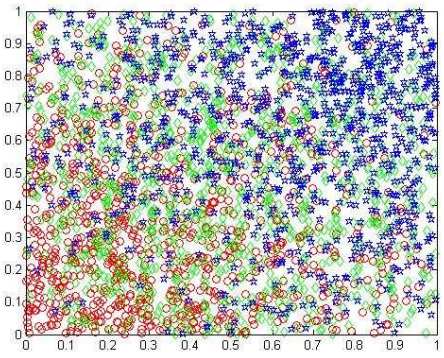
When $p_1 = p_2 = p_3 = 1/3$, the regression function is then an increasing transform of $(x_1, x_2) \in [0,1]^2 \mapsto x_1 + x_2$, given by:

$$\eta(x) = \frac{2.79 \cdot e^{-(x_1+x_2)^2} + 2 \cdot 1.37 \cdot e^{-(x_1+x_2-1)^2} + 3 \cdot 2.79 \cdot e^{-(x_1+x_2-2)^2}}{2.79 \cdot e^{-(x_1+x_2)^2} + 1.37 \cdot \exp^{-(x_1+x_2-1)^2} + 2.79 \cdot e^{-(x_1+x_2-2)^2}}.$$

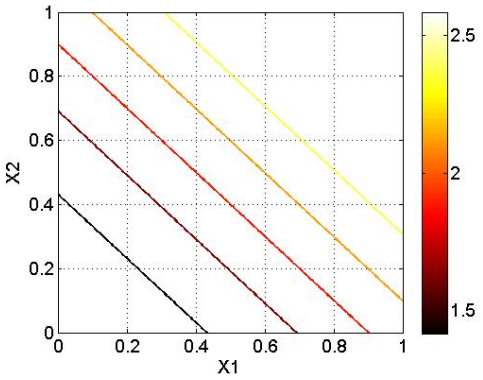
For this distribution we choose $n = 3000$ as the size of a dataset and a simulated dataset is plotted in Fig. 3a, while some level sets of the regression function are represented in 3b.

Mixture of uniform distributions. We consider a mixture of uniform distributions on the unit square$[0,1]^2$, divided into 9 equal parts. The optimal ordering is depicted in Fig. 5.2 b. Table 5.2 displays the values of the regression function $\eta(x)$ and those of the functions $\eta_2(x)/(\eta_1(x) + \eta_2(x))$ and $\eta_3(x)/(\eta_3(x) + \eta_2(x))$ (increasing transforms of $\Phi_{12}(x)$ and $\Phi_{23}(x)$ respectively) on each of the nine cells, showing that **Condition 2.1** is fulfilled. In this case, the size of a dataset is $n = 10000$.

For each one off the distribution, we simulated 50 training samples of size n and a test set of size n. Ranking algorithms have been ran on each training set and the empirical VUS of the resulting scoring rule has been computed on the test set. We



a. Pooled sample: red circles represent instances with label "1", green diamonds those with label "2" and blue stars those with label "3"



b. Optimal level sets.

Figure 3. First example - Mixture of Gaussian distributions

$\phi^*_{1,2}$	$\phi^*_{2,3}$
0.00001	0
0.4000	0
0.8000	0.6000
1.0000	0.8000
1.2500	1.0000
2.5000	1.0000
5.0000	1.6667
∞	2.5000
∞	1000

Table 1. Likelihood ratios

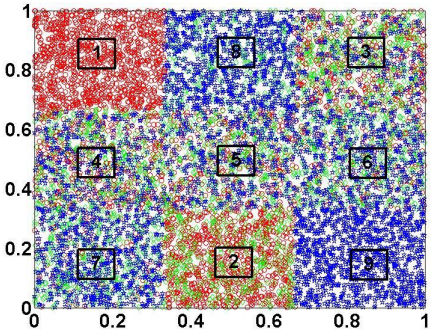


Figure 4. Optimal scoring function s^*

Dataset	TRT	RBpw	SVMl	SVMg	RLSl	RLSg
$\overline{\text{VUS}}$	0.4326	0.4238	0.4334	0.4328	0.4337	0.4330
$\widehat{\sigma}$	0.0073	0.0069	0.0012	0.0036	0.0015	0.0029

Table 2. $\overline{\text{VUS}}$ test (optimal $\text{VUS}^* = 0.4342$)

also calculated the empirical standard deviation over the 50 test ROC surfaces. The results are summarized in Tables 2 and 3. For the gaussian mixture, the TreeRank Tournament is as efficient as the kernels procedures whereas the shape of the level-sets is more difficult to catch with its tree-based structure. For the uniform mixture, the TreeRank Tournament outperforms all the competitors and attains a performance close to the optimal one.

Real datasets. We also applied the TreeRank Tournament algorithm on real data, the *Cardiotocography Data Set* considered in Frank and Asuncion (2010) namely: 2126 fetal cardiotocograms (CTG's in abbreviated form) have been automatically processed and the respective diagnostic features measured. The CTG's have been next analyzed by

Dataset	TRT	RBpw	SVMl	SVMg	RLSl	RLSg
$\overline{\mathrm{VUS}}$	0.5783	0.2984	0.2972	0.3472	0.2969	0.4552
$\widehat{\sigma}$	0.0102	0.0035	0.0004	0.0708	0.0002	0.0018

Table 3. $\overline{\mathrm{VUS}}$ test (optimal $\mathrm{VUS}^* = 0.5926$)

Name	sample size	features space dimension	number of classes
Cardio	2126	20	3
ERA 1-9	1000	4	9
ERA 1-7	951	4	7
ESL 3-7	451	4	9
LEV 0-4	1000	4	5
LEV 0-3	973	4	4
SWD 2-5	1000	10	4
SWD 3-5	978	10	3
MQ2007	69623	46	3
MQ2008	15211	46	3

Table 4. Description of the real datasets

three expert obstetricians and a consensus ordinal label has been then assigned to each of them, depending on the degree of anomaly observed: 1 for "normal", 2 for "suspect" and 3 for "pathologic". We also carried out experiments based on four datasets with ordinal labels (ERA, ESL, LEV and SWD namely), considered in David (2008). Because of the wide disparity between some class sizes, data with certain labels are ignored (in the ESL dataset for instance, the class "1" counts only two observations).

In addition, we considered the LETOR benchmark datasets, available at `research.microsoft.com/en-us/um/people/letor/`. More specifically, we used the two query sets MQ2007 and MQ2008, where pairs "page-query" assigned to a discrete label ranging from 0 to 2 (*i.e.* "non-relevant" - "relevant" - "extremely relevant") are gathered. In both datasets, 46 features are collected, over 69 623 instances in MQ2007 and over 15 211 instances in MQ2008. In these experiments, an estimate of the VUS has been computed through 5 replications of a five-fold cross validation procedure, the results (mean and standard error) are reported in Tables 5. Certain algorithms could not be ran on such datasets, the corresponding programs crashing because of computational difficulties. In this case, "–" is reported in the table. The TreeRank Tournament performs slightly better than its competitors on the LETOR datasets and far outperforms them on the Cardiotocography dataset. Whereas the TreeRank Tournament approach involves estimation of ROC surfaces, most of its competitors require to estimate the regression function. Due to the curse of dimensionality, one may naturally expect that the TreeRank Tournament method performs better (respectively, worse) than regression-based techniques when the number of labels (namely, the dimension of the ROC manifold minus one) is small (respectively large) compared to the dimension of the feature space.

Dataset	TRT	RBpw	SVMl	SVMg	RLSl	RLSg
Cardio	0.8569	0.7165	0.4450	0.2791	0.7788	0.6205
ERA 1-9	0.0023	0.0029	0.0034	0.0020	0.0034	0.0029
ERA 1-7	0.0074	0.0082	0.0088	0.0088	0.0090	0.0080
ESL 3-7	0.6412	0.5745	0.6337	0.6074	0.6387	0.6342
LEV 0-4	0.3003	0.2884	0.3124	0.2847	0.3122	0.3215
LEV 0-3	0.4819	0.4842	0.4968	0.4870	0.4983	0.4954
SWD 2-5	0.4029	0.3304	0.3278	0.3612	0.3316	0.3680
SWD 3-5	0.5674	0.5619	0.5493	0.5599	0.5483	0.5616
MQ2008	0.4115	0.4084	0.4113	–	0.4073	–
MQ2007	0.3172	–	0.2971	–	0.3071	–

Table 5. Comparison of the $\widehat{\mathrm{VUS}}$

6. Conclusion

To the best of our knowledge, the present paper is the first to propose a multipartite ranking algorithm that aims at optimizing *directly* the ROC manifold/surface. As soon as the number K of labels exceeds 3, the challenge arises from the impossibility of interpreting the summary VUS criterion as a cost-sensitive error and multi-class classification algorithms cannot be readily used to optimize local versions of the VUS, in contrast to the bipartite situation. Another difference with the case $K = 2$ lies in the fact that piecewise affine interpolants of the optimal ROC surface are not ROC surfaces in general, making the design of learning algorithms for ROC surface optimization mimicking adaptive approximation schemes very challenging. However, this is precisely what the TreeRank Tournament algorithm introduced in this article achieves: the "Tournament" stage involved in the recursive step of the algorithm permits to extend all the desirable features of the bipartite TreeRank algorithm originally proposed in Clémençon and Vayatis (2009). Beyond theoretical statistical guarantees of the form of rate bounds, the relevance of the TreeRank Tournament algorithm is supported by strong empirical results, displayed here.

Appendix A. VUS formula in the general case

Let $K \geq 2$ and consider independent random variables $X_1, \ldots, X_K$ defined on the same probability space, taking their values in the same space $\mathcal{X}$ and drawn from distributions $F_1, \ldots, F_K$ fulfilling **Assumption 1**. Consider a scoring function $s : \mathcal{X} \to \mathbb{R}$. For any $k \in \{1, \ldots, K-1\}$, set $\mathcal{E}_k(0) = \{s(X_k) < s(X_{k+1})\}$ and $\mathcal{E}_k(1) = \{s(X_k) = s(X_{k+1})\}$. The volume of its ROC manifold is given by:

$$\mathrm{VUS(s)} = \sum_{\mathrm{u} \in \{0,\,1\}^{\mathrm{K}-1}} \frac{\mathbb{P}\{\mathcal{E}_1(\mathrm{u}_1) \cap \cdots \cap \mathcal{E}_{\mathrm{K}-1}(\mathrm{u}_{\mathrm{K}-1})\}}{\mathcal{D}_{\mathrm{u}}}, \tag{A1}$$

where $\mathcal{D}_u = (1+(\tau_1-2)\mathbb{I}\{\tau_1 > 1\}) \times \prod_{j=2}^{K_u} (\tau_j - \tau_{j-1} - 1) \times (1+(K-2-\tau_{K_u})\mathbb{I}\{\tau_{K_u} < K-1\})$, $K_u = K - 1 - \sum_{k=1}^{K-1} u_k$, $\tau_1 = \inf\{k \geq 1 : \ u_k = 0\}$ and $\tau_j = \inf\{k > \tau_{j-1} : \ u_k = 0\}$

for $1 < j \leq K_u$. The TreeRank Tournament algorithm can be straightforwardly extended to the general K-partite ranking setup, considering a tournament between $K-1$ (or $K(K-1)/2$) splitting rule candidates based on the empirical counterpart of (A1).

Appendix B. Technical proofs

Proof of Proposition 4.3

Let $\Delta_J = \max_{0 \leq k < 2^J} \{\alpha^*_{J,k} - \alpha^*_{J,k+1}, \gamma^*_{J,k+1} - \gamma^*_{J,k}\}$. We have:

$$||\mathrm{ROC}^*(.,.) - \mathrm{ROC}(s^*_{2^J},.,.)||_\infty \leq -\frac{\Delta_J^2}{8}\left\{\inf_{(\alpha,\gamma)} \frac{\partial^2}{\partial^2 \alpha}\mathrm{ROC}^*(\alpha,\gamma) + \inf_{(\alpha,\gamma)} \frac{\partial^2}{\partial^2 \gamma}\mathrm{ROC}^*(\alpha,\gamma)\right\}.$$

It thus suffices to establish that $\Delta_J \leq C 2^{-J}$ for some constant $C > 0$. This can be easily established by induction, based on the next lemma. Details are left to the reader.

Lemma B.1: *Let $f : [0,1] \to [0,1]$ be a twice differentiable, decreasing and concave function such that $m_1 \leq f' \leq M_1 < 0$ and $m_2 \leq f" \leq M_2 < 0$.*

(i) Let $x_0 < x_1$ and define x_* such that $f'(x_*) = (f(x_1) - f(x_0))/(x_1 - x_0)$. For $C_2 = 1 - M_2/2m_2$, we have:

$$\max\{x_0 - x_*|, |x_1 - x_*|\} \leq C_2 |x_1 - x_0|.$$

(ii) Let $x_0 < x' < x_1$ such that $\max\{|x_0 - x'|, |x_1 - x'|\} \leq C|x_1 - x_0|$ with $C < 1$. For $C_1 = 1 - (1-C)M_1/m_1$, we have:

$$\max\left\{|f(x_0) - f(x')|, |f(x_1) - f(x')|\right\} \leq C_1 |f(x_1) - f(x_0)|.$$

Proof of Theorem 5.2 (Sketch of)

We shall prove that, for all $\delta \in (0,1)$, we have with probability at least $1-\delta$: $\forall (j,l) \in \{1, \ldots, J\} \times \{0, \ldots, 2^{J-1} - 1\}$,

$$\mathbb{E}\left[\mathbb{I}\{X \in \mathcal{C}^*_{j,2l} \Delta \mathcal{C}_{j,2l}\}\right] \leq C \times B\left((1+a)^j/a^j, n, \delta\right), \tag{B1}$$

for some constant $C < +\infty$, where $B(d,n,\delta) = (c_1^2 V/n)^{1/(2d)} + (c_2^2 \log(1/\delta)/n)^{1/(2d)}$ and Δ denotes the symmetric difference. We start with considering the first iteration. By symmetry, we can assume that $\mathcal{C}_{1,0} = \widetilde{\mathcal{C}}^{(1)}$. Using Lemma 19 in Clémençon and Vayatis (2009), we have, with probability $1-\delta$: $\mathrm{AUC}_{1,2}(s^*_1) - \mathrm{AUC}_{1,2}(s_1) \leq \kappa_1 B(1,n,\delta)$ for some constant $\kappa_1 < +\infty$. By virtue of Lemma 1 in Clémençon et al. (2011), we have

$$0 \leq \mathrm{AUC}_{2,3}(s^*_2) - \mathrm{AUC}_{2,3}(s_2) \leq p_3 p_2/(2(p_3+p_2)) \cdot \mathbb{E}[\mathbb{I}\{X \in \mathcal{C}_{1,0} \Delta \mathcal{C}^*_{1,0}\}].$$

Combining Assumption 5.1 with Hölder inequality, we get that

$$\mathbb{E}\left[\mathbb{I}\{X \in \mathcal{C}_{1,0} \Delta \mathcal{C}^*_{1,0}\}\right] \leq \left(\frac{2(p_1+p_2)}{p_1 p_2}\mathrm{AUC}_{1,2}(s^*_2) - \mathrm{AUC}_{1,2}(s_2)\right)^{\frac{a}{1+a}} \times c^{\frac{1}{1+a}}.$$

Finally, since

$$|\mathrm{VUS}(\mathrm{s}_2^*) - \mathrm{VUS}(\mathrm{s}_2)| \leq |\mathrm{AUC}_{1,2}(\mathrm{s}_2^*) - \mathrm{AUC}_{1,2}(\mathrm{s}_2)| + |\mathrm{AUC}_{2,3}(\mathrm{s}_2^*) - \mathrm{AUC}_{2,3}(\mathrm{s}_2)|$$

(*cf* Theorem 2 in Clémençon et al. (2013)), we have with probability $1-\delta$,

$$|\mathrm{VUS}(\mathrm{s}_2^*) - \mathrm{VUS}(\mathrm{s}_2)| \leq \mathrm{C} \cdot \mathrm{B}((1+\mathrm{a})/\mathrm{a}, \mathrm{n}, \delta).$$

Now let $j > 1$ be fixed and suppose that the bound (B1) holds for $l \leq j-1$. We have

$$\mathrm{VUS}(\mathrm{s}_{2^j}^*) - \mathrm{VUS}(\mathrm{s}_{2^j}) \leq |\mathrm{AUC}_{1,2}(\mathrm{s}_{2^j}^*) - \mathrm{AUC}_{1,2}(\mathrm{s}_{2^j})| + |\mathrm{AUC}_{2,3}(\mathrm{s}_{2^j}^*) - \mathrm{AUC}_{2,3}(\mathrm{s}_{2^j})|.$$

Using the bound established in Clémençon and Vayatis (2009) (see Theorem 15's proof therein), we have

$$\begin{aligned} 2|\mathrm{AUC}_{1,2}(s_{2^j}^*) - \mathrm{AUC}_{1,2}(s_{2^j})| \leq \sum_{l=1}^{2^{j-1}-1} \{ & |F_1(\mathcal{C}_{j-1,l}^*)F_2(\mathcal{C}_{j-1,l}^*)\Lambda_{1,2}(\mathcal{C}_{j,2l}^* \mid \mathcal{C}_{j-1,l}^*) \\ & - F_1(\mathcal{C}_{j-1,l})F_2(\mathcal{C}_{j-1,l})\Lambda_{1,2}(\mathcal{C}_{j,2l} \mid \mathcal{C}_{j-1,l})|\}. \end{aligned}$$

By symmetry, we suppose that the winner of the $(2^{j-1}+l)$-th tournament is $\mathcal{C}_{j,2l} = \arg\max_{\mathcal{C} \in \mathcal{C}_{j-1,l}} \tilde{\Lambda}_{1,2}(\mathcal{C}|\mathcal{C}_{j-1,l})$, *i.e.* the solution of the subproblem 1 *vs* 2. We introduce the set $\bar{\mathcal{C}}_{j,2l} = \arg\max_{\mathcal{C} \subset \mathcal{C}_{j-1,l}} \Lambda_{1,2}(\mathcal{C} \mid \mathcal{C}_{j-1,l})|$. We have

$$\begin{aligned} & |F_1(\mathcal{C}_{j-1,l}^*)F_2(\mathcal{C}_{j-1,l}^*)\Lambda_{1,2}(\mathcal{C}_{j,2l}^* \mid \mathcal{C}_{j-1,l}^*) - F_1(\mathcal{C}_{j-1,l})F_2(\mathcal{C}_{j-1,l})\Lambda_{1,2}(\mathcal{C}_{j,2l} \mid \mathcal{C}_{j-1,l})| \\ & \leq |F_1(\mathcal{C}_{j-1,l}^*)F_2(\mathcal{C}_{j,2l}^*) - F_2(\mathcal{C}_{j-1,l}^*)F_1(\mathcal{C}_{j,2l}^*) - F_1(\mathcal{C}_{j-1,l})F_2(\bar{\mathcal{C}}_{j,2l}) + F_2(\mathcal{C}_{j-1,l})F_1(\bar{\mathcal{C}}_{j,2l})| \\ & + |F_1(\mathcal{C}_{j-1,l})F_2(\bar{\mathcal{C}}_{j,2l}) + F_2(\mathcal{C}_{j-1,l})F_1(\bar{\mathcal{C}}_{j,2l}) - F_1(\mathcal{C}_{j-1,l})F_2(\mathcal{C}_{j,2l}) + F_2(\mathcal{C}_{j-1,l})F_1(\mathcal{C}_{j,2l})| \\ & \stackrel{def}{=} A_{j,2l} + B_{j,2l}. \end{aligned}$$

Using VC inequality just like for the first iteration, we get that, with probability $1-\delta$, the quantity $B_{j,2l}$ is bounded by $B((1+a)/a, n, \delta)$. Notice in particular that we have, with probability $1-\delta$, $\mathbb{E}[\mathbb{I}\{X \in \bar{\mathcal{C}}_{j,2l} \Delta \mathcal{C}_{j,2l}\}] \leq C \cdot B((1+a)/a, n, \delta)$. Reproducing exactly the argument of Clémençon and Vayatis (2009),

$$A_{j,2l} \leq |F_1(\mathcal{C}_{j-1,l}^*) - F_1(\mathcal{C}_{j-1,l})| + |F_2(\mathcal{C}_{j-1,l}^*) - F_2(\mathcal{C}_{j-1,l})| \leq B((1+a)^{j-1}/a^{j-1}, n, \delta)$$

using inequality (B1). Now, observe that

$$\mathbb{E}[\mathbb{I}\{X \in \mathcal{C}_{j,2l}^* \Delta \mathcal{C}_{j,2l}\}] \leq \mathbb{E}[\mathbb{I}\{X \in \mathcal{C}_{j,2l}^* \Delta \bar{\mathcal{C}}_{j,2l}\}] + \mathbb{E}[\mathbb{I}\{X \in \bar{\mathcal{C}}_{j,2l} \Delta \mathcal{C}_{j,2l}\}].$$

Using Hölder inequality and Assumption (5.1), we have

$$\mathbb{E}[\mathbb{I}\{X \in \mathcal{C}_{j,2l}^* \Delta \bar{\mathcal{C}}_{j,2l}\}] \leq C|A_{j,2l}|^{\frac{a}{1+a}} \leq C \cdot B((1+a)^j/a^j, n, \delta).$$

This establishes that inequality (B1) holds for any (j,l) and the desired bound then immediately follows from this (repeating the argument involved at the first iteration).

References

Agarwal, S., Graepel, T., Herbrich, R., Har-Peled, S., and Roth, D. (2005), "Generalization Bounds for the Area Under the ROC Curve," *JMLR*, 6, 393–425.

Clémençon, S., Depecker, M., and Vayatis, N. (2011), "Adaptive partitioning schemes for bipartite ranking," *Machine Learning*, 43, 31–69.

Clémençon, S., Depecker, M., and Vayatis, N. (2012), "An empirical comparison of learning algorithms for nonparametric scoring: the TreeRank algorithm and other methods," *Pattern Analysis and Applications*.

Clémençon, S., Depecker, M., and Vayatis, N. (2013), "Ranking Forests," *Journal of Machine Learning Research*, 14, 39–73.

Clémençon, S., Lugosi, G., and Vayatis., N. (2008), "Ranking and empirical risk minimization of U-statistics.," *Ann. Statist.*, 36, 844–874.

Clémençon, S., Robbiano, S., and N.Vayatis, (2013), "Ranking Data with Ordinal Labels: Optimality and Pairwise Aggregation," *Machine Learning*, 91, 67–104.

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

Clémençon, S., and Vayatis, N. (2010), "Overlaying classifiers: a practical approach for optimal scoring," *Constructive Approximation*, 32, 619–648.

David, A.B., "Ordinal real-world data sets repository," (2008).

Frank, A., and Asuncion, A., "UCI Machine Learning Repository," (2010).

Freund, Y., Iyer, R.D., Schapire, R.E., and Singer, Y. (2003), "An efficient boosting algorithm for combining preferences," *JMLR*, 4, 933–969.

Herbrich, R., Graepel, T., and Obermayer, K. (2000), "Large margin rank boundaries for ordinal regression," *Advances in Large Margin Classifiers*, MIT Press, pp. 115–132.

Hüllermeier, E., Fürnkranz, J., Cheng, W., and Brinker, K. (2008), "Label Ranking by Learning Pairwise Preferences," *Artificial Intelligence*, 172, 1897–1917.

Pahikkala, T., Tsivtsivadze, E., Airola, A., Boberg, J., and Salakoski, T. (2007), "Learning to rank with pairwise regularized least-squares," in *Proceedings of SIGIR*.

Pahikkala, T., Tsivtsivadze, E., Airola, A., Boberg, J., and Salakoski, T. (2007), "Learning to rank with pairwise regularized least-squares," in *Proceedings of SIGIR*, pp. 27–33.

Rajaram, S., and Agarwal, S. (2005), "Generalization Bounds for k-Partite Ranking," in *NIPS 2005 Workshop on Learn to rank*.

Rudin, C., Cortes, C., Mohri, M., and Schapire, R.E. (2005), "Margin-Based Ranking and Boosting Meet in the Middle," in *Proceedings of COLT*.

Scurfield, B. (1996), "Multiple-event forced-choice tasks in the theory of signal detectability," *Journal of Mathematical Psychology*, 40, 253–269.

Waegeman, W., Baets, B.D., and Boullart, L. (2008), "ROC analysis in ordinal regression learning," *Pattern Recognition Letters*, 29, 1 – 9.