Analyzing Ranking Data Using Decision Tree

Philip L.H. Yu, W.M. Wan and Paul H. Lee

Department of Statistics and Actuarial Science, University of Hong Kong

Abstract. Ranking/preference data arises from many applications in marketing, psychology and politics. We establish a new decision tree model for the analysis of ranking data by adopting the concept of classification and regression tree [2]. We modify the existing splitting criteria, Gini and entropy, which can precisely measure the impurity of a set of ranking data. Two types of impurity measures for ranking data are introduced, namely n-wise and top-k measures. Minimal cost-complexity pruning is used to find the optimum-sized tree. In model assessment, the area under the ROC curve (AUC) is applied to evaluate the tree performance. The proposed methodology is implemented to analyze a partial ranking dataset of Inglehart's items collected in the 1993 International Social Science Programme survey. Change in importance of item values with country, age and level of education are identified.

Key words: Decision tree; Ranking data; Impurity function; AUC

1 Introduction

Ranking data are frequently collected when individuals are asked to rank a set of items based on certain pre-defined criterion. It is a simple and efficient way to understand judges' perception and preferences on the ranked alternatives. In many preference studies, ranking responses and additional information about the investigated raters are observed, e.g. socio-economic characteristics. It is often of great interest to determine how these covariates affect the perceived rankings.

Our aim in this paper is to develop new decision tree model to analyze ranking data for discovering the factors that affect the judgement process by which people make choice. It will serve as a complement to existing parametric ranking models (See review in [4], [17] for more details), and algorithms in label ranking and preference learning (See [7], [11] for more details). Decision tree models are nonparametric statistical methodology designed for classification and prediction problems. It produces a set of decision rules for predicting the class of a categorical response variable at the basis of the *input attributes/predictors/covariates*. This classification technique is widely used in statistics, machine learning, pattern recognition and data mining because of its ease of interpretability comparing with other statistical models, and it can handle input attributes in both categorical and interval measurement. Comparing to parametric ranking models, the merit of decision tree lies in its ease of interpretability of nonlinear and interaction effects. Additionally, learning a decision tree can be seen as a process of

variable selection for the data. Questions on adding explanatory variables and interaction terms between variables are handled automatically.

A variety of algorithms have been proposed to construct a decision tree for a single discrete/continuous response in a top-down recursive divide-and-conquer manner, such as ID3 [18], C4.5 [19], CHAID [15] and QUEST [16]. More decision tree algorithms are available in the literature, many of them are a variation of the algorithmic framework mentioned above (See [21] for details). Among all the tree building methodologies, the most popular one is the CART procedure [2]. Construction of CART comprises two stages: *growing* and *pruning*. Detailed review of CART will be provided later in section 2.

Nominal data, ordinal data as well as continuous data can be handled by the decision tree model. It was extended to cope with multivariate data recently through building the tree with a two-stage splitting criteria [22] and through a so-called output kernel trees that are based on a kernelization of the output space of regression trees [8]. Karlaftis [14] used the recursive partitioning models to predict individual mode choice decisions by considering both univariate and multivariate splits. It has been found that trees performed surprisingly well and were comparable to discrete choice logit models. As to the best of our knowledge, modelling ranking data using decision tree has not been studied in literature.

In principle, existing tree models for discrete choice data can be applied to preference data by two approaches. The first approach is to build tree based on the top choice of the given ranking data. Another approach is to treat each ranking of m items as a discrete choice. So each possible ranking outcome contributes to one target level, resulting a total of $m!$ levels. For instance, given three alternatives (a_1,a_2 and a_3), a top-choice tree with 3 target levels or a tree with 6 target levels ($a_1 \succ a_2 \succ a_3$, $a_1 \succ a_3 \succ a_2$, $a_2 \succ a_1 \succ a_3$, $a_2 \succ a_3 \succ a_1$, $a_3 \succ a_1 \succ a_2$ and $a_3 \succ a_2 \succ a_1$) can be constructed.

However, in observational studies, discrete choice tree can provide only limited insights about the underlying behavioral processes that give rise to the data. For the second approach, it will be too heavy-handed in practical, because even moderate values of m would lead to overwhelmingly large number of ranking outcomes ($4! = 24$ and $5! = 120$). Moreover, these nominal trees are restricted only for consistent ranking responses, which all individuals rank the same number of given items. They also are not suitable to handle data with tie ranks because more rank combinations would be involved and this would tremendously increase the number of target levels. Another drawback of this method is ignorance of the ordinal structure in rankings, which is often useful in explaining individuals' preference judgement. Therefore it is impractical to build tree for ranking data using conventional algorithms.

In view of all the limitations and inappropriateness of existing decision tree models for rankings, we are interested to develop a new tree model specifically for preference data. In this article, binary tree is considered. Following the landmark CART procedure, we extend the splitting criteria Gini and entropy to accommodate complete and partial ranking data by utilizing the rank-order structure of preference data in the tree growing process.

Another issue that will be addressed in this paper is the performance assessment of the built decision tree model. The most frequently used performance measure is misclassification rate, which equals to the number of misclassified samples divided by the total number of samples. However, we will not consider it to be the performance measure of our tree model because a sample can either be classified correctly or incorrectly, overlooking the fact that a ranking can be partially agreed with the predicted ranking. That means some items in the rank permutation, but not all, are in the correct ordered position.

We consider goodness-of-fit measures for parametric ranking models for our tree model, such as log-likelihood or other likelihood-based statistics (e.g. BIC). But these approaches may not be suitable because maximizing entropy or deviance is equivalent to maximize the log-likelihood [20]. This will lead to bias towards decision tree that is built on entropy. Therefore, an assessment method independent of the splitting criteria will be more favorable.

The Receiver Operating Characteristic (ROC) curve provides a visualization of the performance of scoring classifier by plotting sensitivity versus 1-specificity at all possible values of the classification threshold. It starts at the bottom-left corner and rises to the top-right corner. Moving along the ROC curve represents trading off false positives for false negatives. In the worst case, random models will run up the diagonal, and the performance of classifier improves as the ROC curve gets near the top-left corner of the plot. Unfortunately, in a comparison of two classifiers, one classifier may not always outperform another at all thresholds. Ambiguous conclusion would be drawn when the two curves intersect. More inadequacies of the ROC curve were discussed in [5].

The area under the ROC curve (AUC) provides a single measure of overall performance of a classifier based on the ROC curve. It is simple and attractive because it is not susceptible to the threshold choice and it is regardless of the costs of the different kinds of misclassification and class priors. The calculation of AUC can be referred to [1] and [9]. The value of AUC always fall within [0.5, 1.0] – it equals 0.5 when the instances are predicted at random and equals 1.0 for perfect accuracy. Statistically, the AUC of a classifier can be seen as the probability that the classifier will rank a randomly chosen positive instance higher than a randomly chosen negative instance. This is equivalent to Mann-Whitney-Wilcoxon test statistic [10].

Traditional ROC curves analysis mainly focus on data with binary target, recently it is extended to multiple class data [9]. In this paper, we adopt the approach of multiclass AUC and generalize the performance measure to pairwise ranking data. The choice of extension to pairwise data over top-2 data is because pairwise data concentrates on two items, while keeping away other irrelevant alternatives.

The remainder of this paper is organized as follows. Section 2 reviews the CART methodology. In section 3, the framework of growing and pruning a decision tree for ranking data is presented. Two new impurity measures, top-k and n-wise measure, are introduced and they are shown to possess the properties of impurity function. Methods for assessing the performance of the tree-structured

classifier are discussed in section 4. To illustrate the feasibility of the proposed algorithm, an example, a simulation study and an application on real data is presented in section 5. Finally, some concluding remarks are given in section 6.

2 Review of Classification and Regression Tree (CART)

Suppose we have a learning sample of size N with measurements $(Y_i, \boldsymbol{X}_i)$, $i = 1, ..., N$, where Y is our target variable and $\boldsymbol{X}$ is the vector of Q predictors $X^q, q = 1, ..., Q$. $\boldsymbol{X}$ and Y can be interval, ordinal or categorical variables. The goal is to predict Y based on $\boldsymbol{X}$ via tree-structured classification.

CART is a binary decision tree that is constructed by recursively partitioning the N learning sample into different subsets, beginning with the root node that contains the whole learning sample. Each subset is represented by a node in tree. In a binary tree structure, all internal nodes have two child nodes whereas the nodes with no descendants are called *terminal/leaf* nodes. At each partition process, a splitting rule $s(t)$, comprises of a splitting variable X^q and a split point, is used to split a group of $N(t)$ cases in node t to left node $N_L(t)$ and right node $N_R(t)$. Decision tree identifies the best split by exhaustive search. The number of possible splits of a categorical predictor X^q of I categories is $2^{I-1} - 1$. For an interval X^q with F distinct values or an ordinal predictor with F ordered categories, $F - 1$ possible splits will be produced on X^q.

2.1 Growing Stage of Decision Tree

The key step of tree growing is to choose a split among all possible splits at each node so that the resulting child nodes are the "purest". To measure the purity of a node t, [2] proposed a measure called impurity function $i(t)$. Let $p(j|t), j \in 1, ..., J$ be the conditional probability of having class j in the learning sample in node t, $\sum_{j=1}^{J} p(j|t) = 1$. Impurity function should satisfy the following three properties: (i) It is minimum when the node is pure ($p(j|t) = 1$ for one $j \in \{1, ..., J\}$); (ii) it is maximum when the node is the most impure ($p(1|t) = ... = p(J|t) = \frac{1}{J}$); (iii) renaming of items doesn't change the node impurity.

It can be shown that if the impurity function is concave, properties 1 and 2 will be satisfied. Property 3 is required because labeling of classes is arbitrary. CART includes various impurity criteria for classification trees, namely the Gini criterion $1 - \sum_{j=1}^{J} p(j|t)^2$ and Twoing criterion. Another frequently used impurity-based criterion applied is entropy $-\sum_{j=1}^{J} p(j|t) log_2 p(j|t)$. Modification of existing measures of node homogeneity is essential for building decision tree model for ranking data and they will be discussed in section 3.1.

Based on the impurity measure for a node, a splitting criterion $\triangle i(s, t)$ can be defined as the reduction in impurity resulting from the split s of node t.

$$\triangle i(s, t) = i(t) - p_L i(t_L) - p_R i(t_R) \tag{1}$$

where $p_L = N_L(t)/N(t)$ and $p_R = N_R(t)/N(t)$ are the proportion of data cases in t to the left child node t_L and to the right child node t_R respectively. The

best split is chosen to maximize a splitting criterion. The concavity property of $i(t)$ assures that further splitting does not increase the impurity, so we can continue growing a tree until every node is pure, and some may contain only one observation. This would lead to a very large tree, that would overfit the data. To eliminate nodes that are overspecialized, pruning is required so that the best pruned subtree can be obtained.

2.2 Pruning Stage of Decision Tree

The minimal cost-complexity pruning method is developed by Breiman *et al.* in 1984 [2]. Before proceeding to the algorithmic framework, some notations are first defined. Let $\tilde{T}$ be the set of terminal nodes of tree T, and the number of terminal nodes, denoted by $|\tilde{T}|$, is defined as the complexity of T. Define $R(t)$ to be the misclassification cost of node t. An obvious candidate of $R(t)$ is the misclassification rate; there are also other choices for the cost function. In a class probability tree, [2] considered pruning with the mean square error which corresponds to take $R(t)$ as the Gini diversity index. For entropy tree, it is natural to take $R(t)$ as deviance. Chou [3] developed a class of divergences in the form of expected loss function and it was shown that Gini, entropy and misclassification rate can be written in the proposed form. In this paper, we specify the cost functions $R(t)$ such that they coincide with impurity functions for ranking data. More details will be given in section 3.3.

For any tree T, the cost-complexity function $R_\alpha(T)$ is formulated as a linear combination of the cost of T and its complexity: $R(T) + \alpha|\tilde{T}|$. The complexity parameter α measures how much additional accuracy a split must add to the entire tree to warrant the addition of one more terminal node. Now consider $T_{t'}$ as the subtree with root t'. As long as $R_\alpha(T_{t'}) < R_\alpha(t')$, the branch $T_{t'}$ contributes less complexity cost to tree T than node t'. This occurs for small α. When α increases to a certain value, the equality of the two cost-complexities is achieved. At this point, the subtree $T_{t'}$ can be removed since it no longer help improving the classification. The strength of the link from node t, $g(t)$, is therefore defined as $\frac{R(t)-R(T_t)}{|\tilde{T}_t|-1}$.

The V-fold cross-validation cost-complexity pruning algorithm works as follows. The full learning dataset L is divided randomly into V equal-size subsets $L_1, L_2, ..., L_V$ and the v^{th} learning sample is denoted to be $L^v = L - L_v$. Using the full learning dataset L, an overly large tree T^0 is built. $g(t)$ are calculated for all internal nodes in T^0 and the node with the minimum value $g(t^1)$ is located. A pruned tree T^1 is created by turning the weakest-linked internal node t^1 into a leaf node. This process is repeated until T^0 is pruned up to the root T^m. Denote α_i be the value of $g(t)$ at the i^{th} stage. A sequence of nested trees $T^0 \supseteq T^1 \supseteq T^2 \supseteq ... \supseteq T^m$ is generated, such that each pruned tree T^i is optimal for $\alpha \in [\alpha_i, \alpha_{i+1})$. Here the word "nested" means that each subsequent tree in the sequence is obtained from its predecessor by cutting one or more subtrees, and thus the accuracy of the sequence of progressively smaller pruned trees decreases monotonically.

Next, for $v = 1, ..., V$, the v^{th} auxiliary maximal tree T_v^0 is constructed based on L^v and the nested sequence of pruned subtrees of T_v^0 is generated ($T_v^0 \supseteq T_v^1 \supseteq T_v^2 \supseteq ... \supseteq T_v^m$). The cross-validation estimate of the misclassification rate $R^{CV}(T^i)$ is then evaluated as $\frac{1}{V}\sum_{v=1}^{V} R(T_v(\sqrt{\alpha_i\,\alpha_{i+1}}))$, where $T_v(\alpha)$ is equal to the pruned subtree T_v^i in the i^{th} stage such that $\alpha_i \leq \alpha \leq \alpha_{i+1}$. Note that the misclassification cost of the pruned subtree $T_v(\sqrt{\alpha_i\,\alpha_{i+1}})$ is estimated by the independent subset L_v. The simplest subtree T^* is selected as the final tree model from $\{T^0, T^1, ..., T^m\}$ by the following rule $R^{CV}(T^*) \leq \min_i R^{CV}(T^i) + SE(R^{CV}(T^i))$. The 1-SE rule is adopted because the position of the minimum $R^{CV}(T^*)$ is uncertain [2]. As a consequence, we get a more conservative estimate for the cross-validated $R^{CV}(T^*)$.

2.3 Class Assignment of Terminal Nodes of Decision Tree

Each terminal node of the final selected tree T^* carries with it a class label $j^* \in \{1, ..., J\}$ which represents the predicted class for target Y of the samples which fall within this node. The class label is usually determined by the plurality rule, so that the misclassification rate of the tree is minimized. Decision tree classifies an instance by passing it down the tree from the root node till it ends up in a class j^* leaf node and obviously the instance will be assigned to class j^*.

3 Decision Tree Model for Ranking Data

In this section, we describe our methodology for constructing decision tree using a learning dataset of rankings. Following the idea of the CART method, our algorithm involves two stages - growing and pruning, to generate the final best subtree. Mathematically, in a completely ranked data of J items, a ranking can be described by a permutation function $\boldsymbol{r} = (r(1), ..., r(J))$ from $\{1, ..., J\}$ onto $\{1, ..., J\}$. The function $r(j)$, $j = 1, ..., J$ is the rank assigned to item j and smaller ranks correspond to the more preferred items.

Let $\boldsymbol{X}$ be a vector of Q covariates $X^q, q = 1, ..., Q$ observed in the data of a preference study and $\boldsymbol{r}$ be the observed ranking responses. We are interested to examine how the covariates affects the N individuals' choice behavior on the basis of the learning sample $(\boldsymbol{r}_l, \boldsymbol{X}_l), l = 1, ..., N$, via tree-based method. Input attributes X can be measured in continuous, ordinal or nominal scale.

3.1 Impurity Measures for Ranking Data

In tree construction, our approach searches for the best splitting rule based on an impurity function. It is not easy to compute the impurity of ranking data based on the permutation function, therefore we introduce two new measures, namely top-k and n-wise measures. Before proceeding, we first define some notations and terminology of choice probabilities and measures for ranking data.

Definition 1. *For top-k measured data* ($k \leq m$) *in node* t, $p_\tau(a_1, ..., a_k | t)$ *and* $N^\tau_{a_1,...,a_k}(t)$ *indicates respectively the proportion and number of judges who rank item* a_1 *first, item* a_2 *the second, and so on, and* a_k *in the* k^{th} *place. Rankings of the remaining* $m - k$ *items are not considered.*

Definition 2. *For n-wise measured data* ($n \leq m$) *in node* t, *let* $p_w(a_1, ..., a_n | t)$ *and* $N^w_{a_1,...,a_k}(t)$ *to be the proportion and number of judges which item* a_1 *ranks higher than* a_2, *which in turn higher than* a_3, *and so on. Items other than* $a_1, a_2, ..., a_n$ *are not taken into consideration.*

For example, in node t, $p_\tau(1, 2 | t)$ denotes the proportion of data which item 1 ranks first, and item 2 ranks second, whereas $p_w(1, 2 | t)$ is the proportion of data which item 1 is preferable to item 2, regardless on the ranks of the two items.

Nevertheless, there are advantages and disadvantages for both methods. The advantage of top-k measure is that existing tree methods for nominal response can be directly applied, owing to the fact that the sum of all proportions of top-k measured data equals one. Therefore impurity measures such as Gini and entropy can still be employed. However, n-wise measured data do not satisfy this property, therefore we need to modify the impurity measures such that they can estimate the node heterogeneity for ranking data based on n-wise comparison. The advantage of n-wise measure is that it takes account of the ordinal nature of ranking. Top-k measures treat every combination of top-k ranking equally, thereby treating preference data as nominal data. For n-wise measure, it models the rankings by making n-wise comparison for all items.

Definition 3. *Given* m *items, denote* $\boldsymbol{\pi}^{m,d}$ *be a set of rankings with members from all* m*-choose-*d *permutations in* $\{1, 2, ..., m\}$. $\boldsymbol{\pi}^{m,d}$ *contains* P^m_d ($C^m_d \times d!$) *rankings coming from* C^m_d *possible ranked* d*-item subsets and each* d*-item subset gives* $d!$ *possible rank permutation. Furthermore, denote a subset of rankings* $\boldsymbol{\pi}^{m,d}_{\{a_1,a_2,...a_d\}}$ *to represent the* $d!$ *rank permutations for item subset* $\{a_1, a_2, ..., a_d\}$ *and* $\boldsymbol{\Omega}^m_d$ *to indicate all* C^m_d d*-item subsets based on* m *items.*

For example, we have an arbitrary item set $\{1, 2, 3, 4\}$, then $\boldsymbol{\pi}^{4,2}$ includes $\{(1, 2)$, $(2, 1)$,$(1, 3)$,$(3, 1)$, $(1, 4)$, $(4, 1)$, $(2, 3)$, $(3, 2)$, $(2, 4)$, $(4, 2)$,$(3, 4)$, $(4, 3)\}$ and all members of $\boldsymbol{\pi}^{4,2}_{\{1,2\}}$ are $\{(1, 2), (2, 1)\}$, whereas $\boldsymbol{\Omega}^4_2$ represents $\{(1, 2)$,$(1, 3)$,$(1, 4)$, $(2, 3)$, $(2, 4)$, $(3, 4)\}$.

3.2 Growing Stage of Decision Tree for Rankings

In section 2 impurity functions for unordered categorical responses are described. Following this reasoning, we provide extension of impurity functions Gini and entropy to deal with ranking data. As mentioned before, top-k ranking data can be viewed as a kind of nominal data, the corresponding impurity functions thus have similar properties with those for nominal target. Properties of n-wise impurity functions are different: (i) it is minimum when there is only one ranking observed for each of C^m_n ranked item subsets; (ii) it attains maximum when all

$n!$ rank permutations are equally distributed in each of C_n^m ranked item subsets; (iii) renaming of items does not change the value of impurity.

Theorem 1 proves that an impurity measure for nominal data can be extended to handle n-wise measured data if it satisfies certain conditions. A definition is given before theorem 1:

Definition 4. *If an impurity function* $i(t) = \phi(p(1|\,t), p(2|\,t), ..., p(J|\,t))$*, satisfying* $\sum_{j=1}^{J} p(j|\,t) = 1$ *can be written as* $\sum_{j=1}^{J} f(p(j|\,t))$*, then it can be generalized to* n*-wise impurity measure, denoted as* $i_w^{(n)}(t) = \phi_w^{(n)}(p_w(r|\,t), \forall r \in \boldsymbol{\pi}^{m,n})$*, with value equals to* $\sum_{r \in \boldsymbol{\pi}^{m,n}} f(p_w(r|\,t))$*.*

Theorem 1. *The* n*-wise impurity function* $i_w^{(n)}(t)$ *satisfies the following conditions:*

1.1 Concave $(\frac{\partial \phi_w^{(n)}(p_w(r|\,t))}{\partial p_w(a|\,t) \partial p_w(b|\,t)}) \leq 0, \forall a, b \in \boldsymbol{\pi}^{m,n}$*.*
1.2 Minimum when one of $p_w(r|\,t), r \in \boldsymbol{\pi}^{m,n}_{\{a_1,...,a_n\}}$ *equals 1* $\forall\ \{a_1, ..., a_n\} \in \boldsymbol{\Omega}_n^m$*.*
1.3 Maximum when all $p_w(r|\,t) = 1/n!$*.*
1.4 Symmetric with respect to $p_w(r|\,t)$*.*

The proof is given in Appendix.

Using n-wise and top-k measures defined above, we can write down, for example, the Gini index of a node t. Given a ranking dataset of m items,

$$\text{top-}k \text{ Gini:} \quad i_\tau^{(k)}(t) = 1 - \sum_{r \in \boldsymbol{\pi}^{m,k}} [p_\tau(r|\,t)]^2 \tag{2}$$

$$n\text{-wise Gini:} \quad i_w^{(n)}(t) = \frac{1}{C_n^m} \sum_{B_n \in \boldsymbol{\Omega}_n^m} \left(1 - \sum_{r \in \boldsymbol{\pi}^{m,n}_{B_n}} [p_w(r|\,t)]^2 \right) \tag{3}$$

The normalizing term $1/C_n^m$ is to bound $i_w^{(n)}(t)$ in the range of 0 and 1. In n-wise impurity measure, $\boldsymbol{\pi}^{m,n}_{B_n}$ denotes the set of permutations for each of all C_n^m ranked item subset in $\boldsymbol{\pi}^{m,n}$. Top-k and n-wise splitting criteria can thus be constructed based on $i_\tau^{(k)}(t)$ and $i_w^{(n)}(t)$ correspondingly to measure the reduction of heterogeneity between two sub-nodes. The split that best separates the parent node into two subgroups having the highest consensus in ranking should be chosen. The node will continue splitting until the node size is less than the user-specified minimum node size value. In our case studies, the minimum node size is set to 1/10 of the training sample size.

3.3 Pruning Stage of Decision Tree for Rankings

We consider pruning the model in a bottom-up manner, using the minimal cost-complexity algorithm introduced in Section 2.2 with 10-fold cross-validation to obtain the final tree that minimizes the misclassification cost. With reference to [3], our cost function $R(t) = P(t) \cdot E_r\left[\ell(r, \hat{p}(r|\,t)) \,|\, t\right]$ is expressed as an expected loss function based on the impurity function of the partition, where $P(t)$ is the

proportion of judges classified into node t in testing data. The loss functions arising from top-k Gini and entropy are

$$\text{top-}k \text{ Gini:} \qquad \ell(r, \hat{p}_\tau(r|\, t)) = 1 + \sum_{r \in \boldsymbol{\pi}^{m,k}} [\hat{p}_\tau(r|\, t)]^2 - 2\hat{p}_\tau(r|\, t) \tag{4}$$

$$\text{top-}k \text{ entropy:} \qquad \ell(r, \hat{p}_\tau(r|\, t)) = -log_2\, \hat{p}_\tau(r|\, t) \tag{5}$$

It should be aware that $\hat{p}_\tau(r|\, t)$ and $\hat{p}_w(r|\, t)$ are evaluated by the learning data, whereas $N_r^\tau(t)$ and $N_r^w(t)$ are obtained from the testing data. The cost function of n-wise impurity measure can be extended analogously, by taking the expectation over all possible C_n^m item subsets.

3.4 Assignment of Terminal Nodes of Decision Tree for Ranking

We consider various approaches to make the assignment. For every leaf node, (i) mean rank of each item is calculated and the predicted ranking is obtained by ordering the mean ranks; (ii) top-choice frequency of each item is calculated and is ordered to give the predicted ranking; (iii) the most frequently observed ranking represents the predicted ranking; (iv) look at the paired comparison probabilities of each item pair or the top-5 most frequently observed ranking responses. The first three approaches reveal the predicted ranking of the items. However, in some situations, the predicted rankings are not of primary concern, when the tree plays a role in facilitating investigation of covariates which influence individuals' difference in item evaluation. For this kind of exploration purpose, method (iv) will give us a more general idea of how the preference orders distributed within a terminal node.

4 Performance Assessment

To compare the performance of the tree models generated by different splitting criteria, we apply the area under the ROC curve in a testing dataset of size N_{ts}. Suppose we have grown a decision tree T with z terminal nodes, the AUC of an item pair (i, j) is calculated as follows:

1. Calculate the pairwise probability $\hat{p}_w(i, j|\, t)$, $t = 1, .., z$ for every leaf node.
2. Assign $\hat{p}_w(i, j|\, t)$ to judges who fall in terminal node t.
3. Rank the judges in the testing dataset in increasing order according to $\hat{p}_w(i, j|\, t)$ and assign rank r_v for the v^{th} individual who prefer item i over item j, $v = 1, ..., N_{ij}^w$. Note that equal rank is assumed when tied.
4. Calculate the number of judges who rank item i higher than item j (c_t), and the number of judges who rank item j higher than item i (d_t) for $t = 1, ..., z$.
5. Compute the sum of the ranks (S) for individuals with preference order $i \succ j$, where $S = \sum_{t=1}^{z} c_t r_t$.
6. Evaluate the AUC of item pair (i, j), A_{ij} by $A_{ij} = \frac{S - c_0(c_0-1)/2}{c_0 d_0}$ where $c_0 = \sum_{t=1}^{z} c_t$ is the total number of judges who rank item i higher than item j, and $d_0 = \sum_{t=1}^{z} d_t$ is the total number of judges who rank item j higher than item i.

The overall performance measure for tree T is defined as the average of AUC over all item pairs $\text{AUC}(T) = \frac{2}{m(m-1)} \sum_{i<j} A_{ij}$ for $i, j = 1, ..., m$, where m is the number of items to be ranked. Tree model with larger AUC reflects better predictive ability. Standard error of the AUC for a two-class problem is given in [9]. However, for multiclass measure of AUC, they recommended using bootstrap method to estimate the standard error because the derivation is difficult.

5 Case Studies

In this section, we illustrate the tree methodology for ranking data described in sections 3 and 4. The first example involves a toy dataset and a simulation ranking data is generated for the second case study. The third application is a real data analysis of political values priority among Europeans.

5.1 Example

A toy example is given to illustrate the performance of different impurity measures. Ranking data is a high dimension data, especially when the number of items is large. Top-k and n-wise measures reduce the dimension of ranking data, and it may lead to information loss. For example, for a ranking dataset of m items, pairwise measure reduces the dataset from $m! - 1$ parameters into C_2^m parameters and top-k measure reduces it into $P_k^m - 1$ parameters. Generally speaking, information loss due to pairwise measure is larger because number of parameters is less. However it may not always be the case.

Suppose we have 32 observations in a ranked dataset of three items. The Gini index of top-3 and pairwise measured data in the parent node t are $i_\tau^{(3)}(t) = 0.8320$ and $i_w^{(2)}(t) = 0.4987$ respectively. Now consider 2 candidate splits by variable A and B that partition the data into the left and right node as below. The

	$N_r^\tau(t)$	$N_r^\tau(t_L)$ in Left Node		$N_r^\tau(t_R)$ in Right Node	
Ranking (r)	in Node t	Split A	Split B	Split A	Split B
1≻2≻3	5	5	5	0	0
1≻3≻2	5	5	0	0	5
2≻1≻3	5	0	5	5	0
2≻3≻1	5	0	0	5	5
3≻1≻2	6	3	3	3	3
3≻2≻1	6	3	3	3	3

Gini reductions of the two splits based on different measures are computed. No difference is observed in the two splits by viewing the data using top-3 measure as both splits give the same Gini reduction of 0.0977.

However, if the impurity reduction is evaluated by pairwise measure, the difference between them will stand out ($\Delta i(A, t) = 0.0977$ and $\Delta i(B, t) = 0.0326$). It is trivial when the preference is presented in paired rankings based on the two splits. Clearly, split based on variable A is preferred. Pairwise measure selects

Paired Ranking (r)	$N_r^w(t_L)$ in Left Node		$N_r^w(t_R)$ in Right Node	
	Split A	Split B	Split A	Split B
$1 \succ 2$	13	8	3	8
$1 \succ 3$	10	10	5	5
$2 \succ 3$	5	10	10	5

the correct splitting variable, where top-3 measure cannot distinguish between the two splits.

5.2 Simulation Study

In this study, pairwise and top-3 measures is compared using a simulation ranking dataset of 3 items. There are two independent variables, namely A and B, which both have two levels 0 and 1. A total of 100 simulation trials has been carried out. In each trial, 40 samples are simulated. Each sample has equal chance of having A to be 0 or 1. If $A = 0$, then the sample must have B to be 0. If $A = 1$, then the chance of having $B = 0$ and $B = 1$ is half half. This results three possible independent variables combinations $(0, 0)$, $(1, 0)$ and $(1, 1)$ with probability 0.5, 0.25 and 0.25.

The ranking responses are generated by ordering the random utility U_i of item i. It is assumed that U_i depends on the two independent variables via $\lambda_{i0} + \lambda_{i1}A + \lambda_{i2}AB + \epsilon_i$ for $i = 1, 2, 3$. Here ϵ_i is a random noise and follows iid $N(0, 1)$. The simulation study is carried out with $(\lambda_{10}, \lambda_{11}, \lambda_{12}, \lambda_{20}, \lambda_{21}, \lambda_{22}, \lambda_{30}, \lambda_{31}, \lambda_{32})$ $= (0, 2, 0, 1, 0, -2, 2, -3, 2)$. In this setting, the corresponding modal rankings of $(A, B) =$ (0,0), (1,0), and (1,1) are $(3 \succ 2 \succ 1)$, $(1 \succ 2 \succ 3)$ and $(1 \succ 3 \succ 2)$ respectively.

It is trivial that the root node should be split according to variable A. For every simulation trial, Gini reduction of the two candidate splits based on pairwise and top-3 measures are calculated to determine which split is preferred. It is found that pairwise measure gives a perfect selection of variable A but top-3 measure mistakenly chooses variable B to split for 19 times. This indicates that pairwise measure performs better in this simulation study. The main reason is that pairwise measure describes a dataset in three parameters $p_w(1, 2|\, t), p_w(1, 3|\, t)$ and $p_w(2, 3|\, t)$, but top-3 measure uses five parameters $p_\tau(1, 2, 3|\, t), p_\tau(1, 3, 2|\, t), p_\tau(2, 1, 3|\, t), p_\tau(2, 3, 1|\, t)$ and $p_\tau(3, 1, 2|\, t)$, and thus the standard error of the impurity is larger for top-3 measured data.

5.3 European Value Priority Data

The partial ranked dataset was obtained from the International Social Service Programme (ISSP) in 1993 [13]. It mainly focused on value orientations, attitudes, beliefs and knowledge concerning nature and environmental issues, and included the so-called Inglehart Index, a collection of four indicators of materialism/ post-materialism as well. Respondents were asked to pick the most important and the second most important goals for their Government from the following four alternatives: (i) Maintain order in nation [ORDER]; (ii) Give people more to say in Government decisions [SAY]; (iii) Fight rising prices [PRICES]

(iv) Protect freedom of speech [SPEECH]. The survey gave a ranked dataset of 5737 observations with top choice and top-2 rankings. In addition, the data provide some judge-specific characteristics and they are applied in tree partitioning. The candidate splitting variables are summarized in Table 1.

Table 1. Description of European ranking data of political values

Covariate	Description / Code	Type	No. of possible values
Country	West Germany=1, East Germany=2, Great Britain=3, Italy=4, Poland=5	Nominal	5
Gender	Male=1, Female=2	Binary	2
Education	0–10 years=1, 11–13 years=2, 14 or more years=3	Ordinal	2
Age	Value ranges from 15 – 91	Interval	76
Religion	Catholic and Greek Catholic=1, Protestant=2, Others=3, None=4	Nominal	4

Respondents can be classified into value priority groups on the basis of their top 2 choices among the four goals. "Materialist" corresponds to individual who gives priority to ORDER and PRICES regardless of the ordering, whereas those who choose SAY and SPEECH will be termed as "post-materialist". The last category is comprised of judges giving all the other combinations of rankings and they will be classified as holding "mixed" value orientations.

Inglehart's thesis of generational based values has been influential in political science since the early 1970s. He has argued that value priorities were shifting profoundly in economically developed Western countries, from concern over sustenance and safety needs toward quality of life and freedom of self-expression, thus from a materialist orientation to a post-materialist orientation [12]. In this analysis, we study the Inglehart hypothesis in five European countries by our decision tree approach, which helps identifying the attributes that affecting Europeans' value priority.

The data is divided randomly into 2 sets, 70% to the learning set for growing the initial tree and finding the best pruned subtree for each of the four splitting criteria; and 30% to the testing set for performance assessment and selection of the splitting criterion to build the final tree. As decision tree is an unstable classifier that small changes in the learning set can cause major changes in the fitted tree structure, we therefore repeat this procedure 50 times and compare the four splitting criteria with their averaged AUC. Lastly, the final tree model is created using the entire dataset for interpretation. Notice that the testing set is not involved in the tree building process and pruned subtree selection, therefore it serves as an out-of-sample dataset for model comparison. The four splitting criteria for rankings include top-2 and pairwise measure of Gini and entropy. Here, we apply pairwise and top-2 measures as the data only contain individuals' preference orders of the most and the second most desirable goals. Table 2 shows the averaged AUC and their standard error of the best pruned subtrees for each splitting criterion based on 50 repetitions. The tree structure and performance of the final models are also presented in the same table. Figure

1 displays the six ROC curves of each item pairs arise from the top-2 entropy tree. The tree did a better job of predicting the item pair "SAY vs PRICES", but poor for "SAY vs SPEECH". We do not illustrate the ROC curves of other trees as the performance of the four trees are comparable and it is hard to distinguish them in the graph.

Table 2. Summary of the best pruned subtrees of 4 splitting criteria

Method	Avg. AUC	S.E	AUC	No. of Leaves	Depth
Top-2 entropy	0.61947	0.0056	0.62951	12	5
Pairwise Gini	0.61896	0.0058	0.62902	12	5
Pairwise entropy	0.61857	0.0056	0.62709	11	5
Top-2 Gini	0.61425	0.0063	0.61931	9	4

The four tree models are found to have similar node partitions. The root node is split according to whether the judges came from Poland or not (country $= 5$ vs $\neq 5$). At the second level, the splits are based on age. For Polish, the respondents are divided with the rule "age<59?", while the remaining judges are split according to age<53 or not. Further partitions involved education level, country and age. The factors religion and gender seem not to be influential. It is observed that in the learning phase, top-2 Gini tends to give a smaller tree while top-2 entropy gives a more complicated tree on average. Based on the assessment criterion, the top-2 entropy tree is chosen as the best model and it is applied for further analysis. As shown in Figure 2, this tree has 5 levels of depth and 12 leaves. For sake of brevity, we do not show the other three tree structures. Summary of the terminal nodes of the final tree is reported in two tables. Table 3 lists the mean rank of the four political goals and the three most frequent top-2 ranking, whereas Table 4 shows the individuals' value priority and the proportion of six pairs of political goals in each leaf node.

We now turn to examine the covariate and interaction effects based on the final tree model. In Poland, individuals were more likely to favor materialistic items ORDER and PRICES (in leaves 5, 8 and 9). In East Germany, judges appeared to support ORDER and SAY more, particularly those older generations gave higher priority to ORDER (in leaf 12). Respondents of West Germany showed stronger emphasis on SAY. Those better educated West Germans were more postmaterialist than the lower educated ones as they preferred SAY and SPEECH, rather than the other two materialist items (in leaf 15). Mixed value orientations were anchored in British because all the related leaf nodes give us a preference prediction of ORDER $\succ$ SAY or SAY $\succ$ ORDER.

The result can be summed up in two observations: (i) Despite some cross-national differences, our findings do not deviate much from Inglehart's theory, which claimed that societies embrace post-materialist values as they move towards more economic security and affluence. The older European generations experienced economic and social insecurity in their preadult years during World War II, they thus gave stronger concern on the materialist values compared to the younger cohorts. Younger post-war generations developed post-materialist

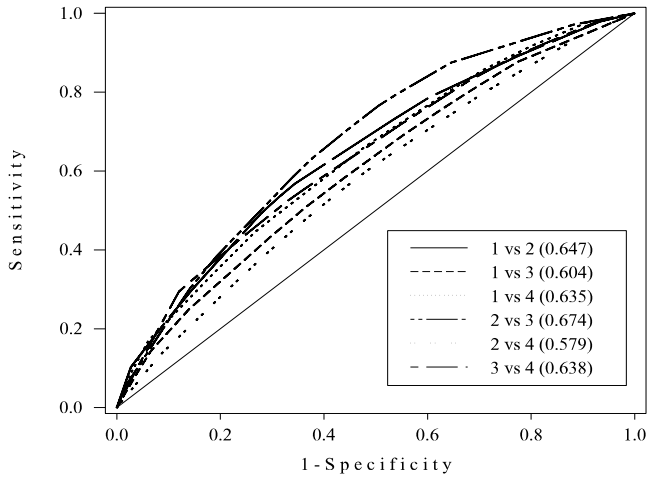

Fig. 1. ROC curves of top-2 entropy tree. The four value items are coded as follows 1=[ORDER], 2=[SAY], 3=[PRICES] and 4=[SPEECH]. The 45° diagonal line connecting (0,0) and (1,1) is the ROC curve corresponding to random chance. Given next to the legends are the areas under the corresponding dashed ROC curves.

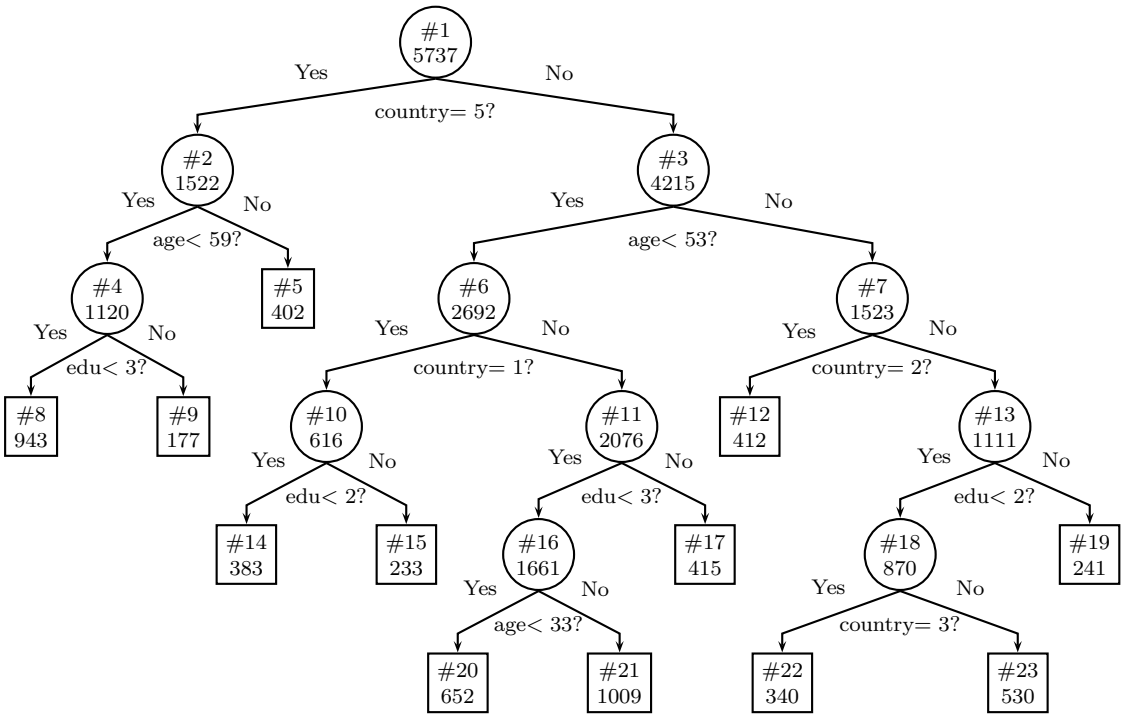

Fig. 2. Tree structure diagram based on top-2 entropy. In each node, the node ID and the number of judges are shown. The splitting rule is given under the node. The abbreviation "edu" stands for the variable education.

Table 3. Importance of 4 political values in terminal nodes of top-2 entropy tree

	Node	Mean rank				Frequent top-2 ranking		
Node(t)	Size	ORDER	SAY	PRICES	SPEECH	1st	2nd	3rd
5	402	1.69	3.08	1.99	3.24	1,3 (40.3%)	3,1 (24.9%)	1,2 (7.5%)
8	943	2.10	2.70	1.95	3.26	3,1 (25.7%)	1,3 (22.5%)	3,2 (13.0%)
9	177	2.16	2.49	2.40	2.95	1,3 (17.3%)	3,1 (13.7%)	1,2 (12.2%)
12	412	1.65	2.42	2.63	3.30	1,3 (27.5%)	1,2 (22.2%)	2,1 (18.2%)
14	383	2.33	2.11	2.64	2.92	2,1 (17.0%)	2,3 (12.3%)	2,4 (12.0%)
15	233	2.73	1.86	2.97	2.44	2,4 (25.9%)	4,2 (15.7%)	2,3 (11.7%)
17	415	2.31	2.05	2.80	2.83	2,4 (15.7%)	2,1 (14.9%)	1,2 (14.2%)
19	241	1.88	2.40	2.83	2.89	1,3 (22.1%)	2,1 (14.9%)	1,2 (13.8%)
20	652	2.28	1.95	2.67	3.09	2,1 (20.1%)	2,3 (19.1%)	1,2 (13.4%)
21	1009	2.09	2.20	2.62	3.10	1,3 (18.0%)	1,2 (16.1%)	2,1 (15.8%)
22	340	2.22	2.30	2.40	3.07	1,3 (18.1%)	2,3 (15.9%)	1,2 (12.3%)
23	530	1.83	2.62	2.48	3.07	1,3 (28.2%)	1,2 (16.2%)	3,1 (9.6%)

Remark: In the last three columns, the code 1 - 4 represents each of the political goals: 1=[ORDER], 2=[SAY], 3=[PRICES] and 4=[SPEECH]; "i, j" implies goal $i \succ$ goal j and the percentage beside indicates the proportion of instances having the corresponding top-2 ranking in node t.

values as they grew up during periods of relative prosperity. (ii) There is a clear tendency in each country for the higher educated to be the more postmaterialist groups. Duch and Taylor [6] stated that the post-materialist items tap certain fundamental democratic values, such as liberty and rights consciousness. The better educated would have had more opportunity to learn to appreciate such principles, and thus they will prefer post-materialist items more.

For comparison, we tried to learn a decision tree in another setting for this dataset, by transforming the top-2 ranking problem into six binary classification problems of pairwise preferences ($1 \succ 2$ vs $2 \succ 1, \ldots, 3 \succ 4$ vs $4 \succ 3$). However, due to large proportion of ties in some pairwise preferences (61.6% for {2,4} and 71.5% for {3,4}), not all information can be utilized to build this alternative tree model and so model comparison is not relevant.

6 Conclusion

We have investigated the use of decision tree model for analyzing ranking data, which makes explanation of individuals' rank-order preference differences easier compared to existing parametric ranking models and algorithms in label ranking and preference learning, especially when non-linearity and high-order interactions are involved in the studied covariates. It is noteworthy that our tree methodology includes the multinomial tree as a special case and it can accommodate inconsistent rankings, as well as tie rankings. We have proposed two impurity measures, namely n-wise and top-k measures, to evaluate the goodness of split for ranking data. Examples and simulations showed that the established impurity functions effectively measure the node heterogeneity. It is interesting

Table 4. Value priority and pairwise probabilities in leaf nodes of top-2 entropy tree

Node			Pairwise probabilities					
Node(t)	Size	Value	$p_w(1,2\vert\, t)$	$p_w(1,3\vert\, t)$	$p_w(1,4\vert\, t)$	$p_w(2,3\vert\, t)$	$p_w(2,4\vert\, t)$	$p_w(3,4\vert\, t)$
5	402	M	83.1%	60.8%	87.3%	21.0%	53.7%	83.1%
8	943	M	65.3%	45.9%	79.3%	31.0%	64.8%	82.1%
9	177	M	58.8%	55.9%	69.2%	47.7%	61.9%	63.8%
12	412	B	67.5%	77.7%	90.0%	53.9%	72.0%	68.3%
14	383	B	27.4%	32.1%	41.0%	37.6%	40.1%	36.0%
15	233	P	44.1%	57.8%	64.8%	63.6%	69.5%	57.3%
17	415	B	44.5%	63.1%	61.0%	68.3%	71.1%	51.0%
19	241	B	60.8%	76.3%	74.5%	58.3%	62.2%	51.9%
20	652	B	41.6%	59.7%	70.6%	68.9%	77.3%	61.5%
21	1009	B	52.6%	64.5%	74.1%	60.0%	72.8%	62.6%
22	340	B	51.3%	55.6%	71.2%	52.8%	68.2%	67.9%
23	530	M	69.0%	69.1%	78.7%	45.7%	61.8%	66.8%

Remark: The third column "Value" shows the value priority group of judges in each leaf node, where B=Mixed values; M=Materialist and P=Post-materialist. For column 4 to 9, the four political goals are labeled as: 1=[ORDER], 2=[SAY], 3=[PRICES] and 4=[SPEECH].

to find that pairwise impurity measure in some instances is more preferred than top-k measure. The main reason is that pairwise measure describes a dataset with less parameters, and thus the corresponding standard error of the impurity is smaller.

The tree algorithm is easy to implement and flexible that we can specify the number of ranks used in splitting for the top-k measures and the value of n in the n-wise measure according to the ranking data being analyzed. To assess the predictive performance of the final tree, the AUC is used for the purpose. In the real application, the AUCs of the four competitive trees are compared. It is important to emphasize that we are not trying to draw any conclusion about which splitting criterion is more superior, as it is definitely related to the types of the observed rankings.

Acknowledgments

The research of Philip L. H. Yu was supported by a grant from the Research Grants Council of the Hong Kong Special Administrative Region, China (Project No. HKU 7473/05H).

Appendix

In this appendix, the proof of Theorem 1 will be provided. Recall that the size of the ranked item set is m. We hereafter omit t which stands for node t to simplify the notation of proportion and impurity.

Proof of Theorem 1

Proof. Denote $p_w(r), \forall r \in \boldsymbol{\pi}^{m,n}$ as $\boldsymbol{p}_\gamma$, and thus we have $i_w^{(n)} = \phi_w^{(n)}(\boldsymbol{p}_\gamma)$. Also, define B_n to be a n-item subset from $\{1, ..., m\}$

1.1 Let $a, b \in \boldsymbol{\pi}^{m,n}$. If $a \neq b$, it is trivial that $(\frac{\partial \phi_w^{(n)}(\boldsymbol{p}_\gamma)}{\partial p_w(a) \partial p_w(b)}) = 0$.

If $a = b$, since $\phi_w^{(n)}(\boldsymbol{p}_\gamma)$ can be written as

$$\sum_{B_n \in \boldsymbol{\Omega}_n^m} \sum_{r \in \boldsymbol{\pi}_{B_n}^{m,n}} f(p_w(r)) \tag{6}$$

and $\sum_{r \in \boldsymbol{\pi}_{B_n}^{m,n}} p_w(r) = 1$, hence $f(p_w(r))$ is concave, and the permutation a (WLOG assume $a = \{a_1, ..., a_n\}$) will only appear once in the function $\phi_w^{(n)}(\boldsymbol{p}_\gamma)$, therefore $(\frac{\partial \phi_w^{(n)}(\boldsymbol{p}_\gamma)}{\partial p_w(a) \partial p_w(b)}) = (\frac{\partial f(p_w(a))}{\partial p_w(a) \partial p_w(a)}) \leq 0$ (sum of concave functions is concave).

1.2 Since $\phi_w^{(n)}(\boldsymbol{p}_\gamma)$ can be written as equation (6), and note that $\sum_{r \in \boldsymbol{\pi}_{B_n}^{m,n}} p_w(r) = 1, \forall B_n \in \boldsymbol{\Omega}_n^m$, therefore minimizing $\phi_w^{(n)}(\boldsymbol{p}_\gamma)$ will be equivalent to minimize $\sum_{r \in \boldsymbol{\pi}_{B_n}^{m,n}} f(p_w(r)), \forall B_n \in \boldsymbol{\Omega}_n^m \in \{1, ..., m\}$. The condition will be for each of $B_n \in \boldsymbol{\Omega}_n^m$, one of the ranking probabilities $p_w(r), r \in \boldsymbol{\pi}_{B_n}^{m,n}$ equals 1.

Note that n-wise measured data are derived from full ranking, and some combinations of n-wise data are intransitive. For example, it is impossible to have $p_w(1,2) = 1$, $p_w(2,3) = 1$ and $p_w(3,1) = 1$. Another contradictory example is $p_w(1,2,3) = 1$, $p_w(3,2,4) = 1$. Therefore, by eliminating those intransitive n-wise data combinations, the minimizing condition of n-wise impurity functions can be reduced to: one of the probability $P(a_1 \succ a_2 \succ ... \succ a_m) = 1$ and all other full ranking probabilities equal to zero.

1.3 Since $\phi_w^{(n)}(\boldsymbol{p}_\gamma)$ can be written as equation (6), and note that $\sum_{r \in \boldsymbol{\pi}_{B_n}^{m,n}} p_w(r) = 1, \forall B_n \in \boldsymbol{\Omega}_n^m$, therefore maximizing $\phi_w^{(n)}(\boldsymbol{p}_\gamma)$ will be equivalent to maximize $\sum_{r \in \boldsymbol{\pi}_{B_n}^{m,n}} f(p_w(r)), \forall B_n \in \boldsymbol{\Omega}_n^m \in \{1, ..., m\}$. The condition will be for each of $B_n \in \boldsymbol{\Omega}_n^m$, all ranking probabilities $p_w(r), r \in \boldsymbol{\pi}_{B_n}^{m,n}$ equal to $1/n!$.

The n-wise impurity function is maximized at all $p_w(r)$ equal. However, unlike the above minimum case, it cannot be generalized to the case when all full ranking probabilities are uniformly distributed. For example, a full ranking of 3 items with $p_\tau(1,2,3) = p_\tau(3,2,1) = 0.5$ implied pairwise measure of $p_w(1,2) = p_w(2,1) = p_w(1,3) = p_w(3,1) = p_w(2,3) = p_w(3,2) = 0.5$. However, under special condition, $p_w(r) \forall r \in \boldsymbol{\pi}^{m,n}$ equal implies evenly-distributed full ranking probabilities.

1.4 When item a_i and a_j are swapped, all values of $p_w(r)$ with $a_i, a_j \in r$ and $p_w(r)$ with $a_i, a_j \notin r$ remain the same. All values of $p_w(r)$ with $a_i \in r, a_j \notin r$ and $p_w(r)$ with $a_j \in r, a_i \notin r$ exchange. Afterall, there is no effect in $i_w^{(n)}$.

References

1. Bradley, A.P.: The use of the area under the ROC curve in the evaluation of machine learning algorithms. Pattern Recognition 30, 1145-1159 (1997)
2. Breiman, L., Friedman, J.H., Olshen, R.A., Stone C.J.: *Classification and Regression Trees*. Belmont, California: Wadsworth (1984)
3. Chou, P.A.: Optimal partitioning for classification and regression trees. IEEE Transactions on Pattern Analysis and Machine Intelligence 13, 340–354 (1991)
4. Critchlow, D.E., Fligner, M.A., Verducci, J.S.: Probability models on rankings. Journal of Mathematical Psychology 35, 294–318 (1991)
5. Drummond, C., Holte, R.C.: What ROC curves can't do (and cost curves can). In: Proceedings of the 1st Workshop on ROC Analysis in AI, pp. 19-26. Valencia, Spain (2004)
6. Duch, R.M., Taylor, M.: Postmaterialism and the economic condition. American Journal of Political Science 37, 747–778 (1993)
7. Fürnkranz, J., Hüllermeier, E.: Pairwise preference learning and ranking. In: Proceedings of the 14th European Conference on Machine Learning (ECML-03), pp. 145-156. Cavtat, Croatia (2003). Springer-Verlag.
8. Geurts, P., Wehenkel, L., Florence, A.: Kernelizing the output of tree-based methods. In: Proceedings of the 23rd International Conference on Machine Learning (ICML-06), pp. 345–352. Pittsburgh, Pennsylvania (2006).
9. Hand, D.J., Till, R.J.: A simple generalisation of the area under the ROC curve for multiple class classification problems. Machine Learning 45(2), 171–186 (2001)
10. Hanley, J.A., McNeil, B.J.: The meaning and use of the area under a receiver operating characteristic (ROC) curve. Radiology 143, 29-36 (1982)
11. Hüllermeier, E., Fürnkranz, J.: On Minimizing the Position Error in Label Ranking. In: Proceedings of the 17th European Conference on Machine Learning (ECML-07), pp. 583–590. Warsawa, Poland (2007). Springer-Verlag.
12. Inglehart, R.: *The Silent Revolution: Changing Values and Political Styles among Western Publics*. Princeton Univerity Press, Princeton (1977)
13. Jowell, R., Brook, L., Dowds, L.: *International Social Attributes: the 10th BSA Report*. Dartmouth Publishing, Aldershot (1993)
14. Karlaftis, M.: Predicting mode choice through multivariate recursive partitioning. Journal of transportation engineering 130(22), 245–250 (2004)
15. Kass, G.V.: An exploratory technique for investigation large quantities of categorical data. Applied Statistics 29, 119–127 (1980)
16. Loh, W.Y. Shih, Y.S.: Split selection methods for classification trees. Statistica Sinica 7, 815–840 (1997)
17. Marden, J.I.: *Analyzing and Modeling Rank Data*. Chapman and Hall (1995)
18. Quinlan, J.R.: Induction of decision trees. Machine Learning 1, 81–106 (1986)
19. Quinlan, J.R.: C4.5: *Programs for Machine Learning*. Morgan Kaufmann Publishers (1993)
20. Ripley, B.D.: *Pattern Recognition and Neural Networks*. Cambridge University Press, Cambridge (1996)
21. Rokach, L., Maimon, O.: Decision trees. In: Maimon, O., Rokach, L. (eds.), *The Data Mining and Knowledge Discovery Handbook*, pp. 165–192. Springer, Berlin (2005)
22. Siciliano, R., Mola, F.: Multivariate data analysis and modeling through classification and regression trees. Computational Statistics and Data Analysis 32, 285–301 (2000)