Growing Decision Trees in an Ordinal Setting

Kim Cao-Van,[†] Bernard De Baets[*]
Department of Applied Mathematics, Biometrics, and Process Control, Ghent University, Coupure links 653, B-9000 Gent, Belgium

Although ranking (ordinal classification/regression) based on criteria is related closely to classification based on attributes, the development of methods for learning a ranking on the basis of data is lagging far behind that for learning a classification. Most of the work being done focuses on maintaining monotonicity (sometimes even only on the training set). We argue that in doing so, an essential aspect is mostly disregarded, namely, the importance of the role of the decision maker who decides about the acceptability of the generated rule base. Certainly, in ranking problems, there are more factors besides accuracy that play an important role. In this article, we turn to the field of multicriteria decision aid (MCDA) in order to cope with the aforementioned problems. We show that by a proper definition of the notion of partial dominance, it is possible to avoid the counter-intuitive outcomes of classification algorithms when applied to ranking problems. We focus on tree-based approaches and explain how the tree expansion can be guided by the principle of partial dominance preservation, and how the resulting rule base can be graphically represented and further refined. © 2003 Wiley Periodicals, Inc.

1. INTRODUCTION

Monotone (or positive) classification, ordinal classification, ordinal regression, and ranking (or sorting) are all terms describing essentially one and the same problem: learning an ordinal[a] function, which generally should be monotone. The first two arose from machine learning, the third one from statistics, and the last one from the multicriteria decision aid (MCDA), including preference modeling. These three research communities have their own agenda and follow more or less independent paths of investigation, thus, the mentioned variety of terms. It also shows the wide interest in this type of problem, which frequently arises in social

[*]Author to whom all correspondence should be addressed: e-mail: Bernard.DeBaets@ugent.be.
[†]e-mail: Kim.CaoVan@ugent.be.

[a]A discrete measurement scale in which only the order counts, i.e., intervals are meaningless.

INTERNATIONAL JOURNAL OF INTELLIGENT SYSTEMS, VOL. 18, 733–750 (2003)
 Published online in Wiley InterScience (www.interscience.wiley.com). • DOI 10.1002/int.10113

sciences, medicine, risk analysis, information retrieval with human preferences, etc.

All of these approaches have in common the fact that they deal in one way or another with some kind of monotonicity. The ordinal learning model (OLM) of Ben-David[1,2] builds a monotone rule base starting from a not necessarily monotone training set. OLM is an instance-based classifier: the rule base consists of examples and a prescribed manner on how to deal with new examples based on the examples in the rule base. Nonmonotonicity is resolved by the principle of conservatism. Although OLM has proven to give good results, example-based rule bases are hard to interpret. Moreover, sometimes the algorithm performs calculations that are not meaningful on an ordinal scale.

A monotone extension of ID3 (MID) was proposed later by the same author[3] using another impurity measure for splitting, the total ambiguity score. However, the resulting tree may not be monotone anymore even when starting from a monotone data set.

Makino et al.[4] proposed a monotone (or positive) decision tree (P-DT) and a quasi-monotone (quasi-positive) decision tree (QP-DT) extension of ID3 in the two-class setting. They start from a monotone training set and demand, in the case of QP-DT, that monotonicity is (only) guaranteed on this training set, while in the case of P-DT the tree (or equivalently, the derived rule base) is required to be monotone. These methods have been nontrivially extended by Potharst[5,6] to the k-class problem. Moreover, he also accommodates continuous attributes. In addition to the fact that these approaches start from a monotone training set, the main technique for guaranteeing (quasi-)monotonicity is by adding at each step, if necessary, new data generated from the data in the previous step. Although some techniques for adding data have already proven useful for classification purposes,[7] it is not clear what the impact is of the techniques used in the aforementioned studies in the case of ranking. We feel uncomfortable with the fact that the addition of nonauthentic data forms an essential part of an algorithm.

Greco et al.[8,9] proposed the dominance-based rough set (DBRS) approach as an extension of the classical rough set approach. They are guided by principles of monotonicity derived from preference modeling (the outranking approach), which leads to defining upper and lower approximations based on the dominance relation. The result is a rule base that reflects the degree of monotonicity of the training data, with a quasi-monotone rule base if the training set was monotone. The consequences of these rules are expressed in terms of inequalities rather than equalities as in the previous methods. Some concerns have been raised and improvements have been suggested by Düntsch and Gediga.[10] Recently, Bioch and Popova[4] have introduced an alternative extension based on the concepts of the monotone indiscernability matrix and monotone reduct but the theory was developed starting from a monotone data set.

Last, Herbrich[12,13] developed a distribution independent model of ordinal regression, influenced by notions from the field of preference modeling (the utility approach). Then, this model is used to construct a support vector machine. However, the type of monotonicity dealt with is not the same as in the problem of

ranking. Nevertheless, most of the ideas elaborated are more than interesting in the present context.

All of the aforementioned methods (except DBRS) reside in the machine learning paradigm because their main concern is to improve on prediction accuracy. This is, of course, a noble goal, but in doing so, they tend to neglect the human factor in many applications whether a certain rule base is used not only depends on its accuracy, but also on its acceptability. This means that the rule base should be intuitive (which implies that monotonicity should be required, not only quasi-monotonicity), rich in information, and as easy as possible to read. For example, univariate (or axis-parallel) tree structures are known to have an easy interpretation; however, they do not render the (non)presence of monotonicity transparent and hence are not the most suitable representation for ranking problems.

Our philosophy is nested in MCDA; our main concern is to derive a rule base (in this study it also can be represented by a tree because of the greedy top-down induction) that is acceptable to a decision maker as discussed previously. We realize this by extending the notion of partial dominance relation introduced in Ref. 8 and providing it with proper semantics. This enables us to introduce a less restrictive form of monotonicity, captured by the *principle of partial dominance preservation.* Another important feature of our approach is that the leaves of our tree are labeled with intervals. The natural occurrence of intervals in the problem of ranking is not new to MCDA and also was mentioned in Ref. 6 but not recognized to its full extent. In addition, it was implicitly used in Ref. 9.

It is not our intention here to give a ready-to-use algorithm for learning a ranking. We believe it is more enriching to focus on one (in this case probably the most important) aspect of such an algorithm than to drown the reader in a flood of technicalities resulting in an *input-output machine* where many of the internal processes are difficult to understand or interpret. Even the basic notions underlying the various aspects of such an algorithm are full of important subtleties making it impossible (or at least not advisable) to treat them all at once.

2. CLASSIFICATION AND RANKING

Mostly, a classification is defined directly over a measurement space X (see, e.g., Refs. 14 and 15). However, we would like to make a distinction between the object space Ω and its operationalization, the measurement space X.

DEFINITION 1. *A classification* (*rule*) *in* Ω *is the assignment of the objects belonging to* Ω *to some element called a class label, in a universe* $\mathcal{D}$. *We will assume* $\mathcal{D}$ *to be finite. In other words, a classification* (*rule*) *is some mapping*

$$f : \Omega \rightarrow \mathcal{D}$$

The class labels can be identified with their inverse image in the object space Ω, *where they constitute a partition. We will call these inverse images* (*object*) *classes.*

Hence, the set of all classifications in Ω is in one-to-one correspondence with the set of all partitions of Ω.

In the context of learning a classification, where a whole classification has to be induced from a partially given one, the previous definition still is not satisfactory because the object space Ω bears no structure and therefore no induction can be performed. For this reason, the objects are represented by means of some of their properties, captured by a finite set of attributes $Q = \{q_1, \ldots, q_n\}$, where q_i is a mapping from Ω to some value set X_{q_i}. The value sets X_q can be ordered (numerical), e.g., size, or unordered (symbolic), e.g., color. Any object can now be identified with a point or vector in the measurement space $X = X_{q_1} \times \ldots \times X_{q_n}$, which hence constitutes an operationalization of the object space.

Once this representation has been done, there exists many methods and algorithms to perform the induction on the basis of a sample (training set). However, all of these methods fail when applied to a problem very similar to classification, namely, ranking.

A ranking is like a classification except that it has a semantic structure on the classes, which bear labels such as {bad, moderate, good} or {low, middle, high}. This additional structure can be expressed easily by a complete order[c] on the decision space $\mathcal{D}$.

DEFINITION 2. *A ranking in Ω is a classification $f : \Omega \to \mathcal{D}$, together with a complete order $\leq_d$ on $\mathcal{D}$, i.e., $(\mathcal{D}, \leq_d)$ is a chain. We denote this ranking by $(f, \leq_d)$. Moreover, the order $\leq_d$ defines a weak preference relation S on Ω as follows:*

$$(\forall\, r \in \mathcal{D})(\forall\, s \in \mathcal{D})(\forall\, a \in C_r)(\forall\, b \in C_s)(aSb \Leftrightarrow r \geq_d s)$$

where $C_d = f^{-1}(d)$ is the class associated with the class label $d \in \mathcal{D}$.

The relation S is reflexive and aSb is interpreted as "a is at least as good as b." In this way, we have a preference structure on Ω linked with the classes (see Ref. 16). It is well known that this relation can be split into a strict preference relation P and an indifference relation I, with $S = P \cup I$. Let $a, b \in \Omega$, assume $a \in C_r$ and $b \in C_s$ with $r, s \in \mathcal{D}$, and then we have

$$aPb \Leftrightarrow r >_d s \quad \text{and} \quad aIb \Leftrightarrow r = s.$$

As in the case of classifications, it is necessary to represent the objects. Here, however, attributes no longer suffice, and we also have to use criteria (see Ref. 16).

DEFINITION 3. *A criterion is a mapping $c : \Omega \to (X_c, \leq_c)$, where $(X_c, \leq_c)$ is a chain, such that it appears meaningful to compare two objects a and b, according to a particular point of view, on the sole basis of their evaluations $c(a)$ and $c(b)$.*

In essence, a criterion is an attribute that is accompanied by an order $\leq_c$ that relates to the order $\leq_d$ on the classes. In this study, we restrict ourselves to so-called true criteria,[5] for which it holds that "object a is at most as good as object b according

[c]A complete order on a set Ω is a reflexive, antisymmetric, transitive, and complete binary relation on Ω.

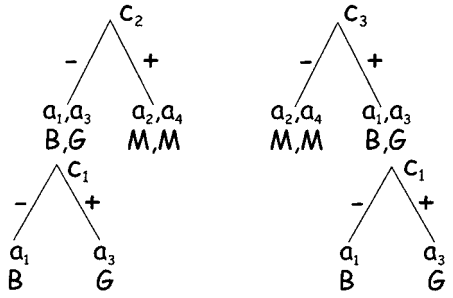

Figure 1. Classification trees T_1 and T_2.

to criterion c if and only if $c(a) \leq_c c(b)$." The objects still can be represented in a measurement space X on the basis of a finite set of criteria $C = \{c_1, \ldots, c_n\}$. Hence, it also is possible to consider the product order $\leq_X$ on X.

3. THE PRINCIPLE OF DOMINANCE PRESERVATION

Example (Part 1). Assume four candidates are applying for a job. They are evaluated according to their working experience (little or much), their capacity for learning (slow or fast), and their personal profile, i.e., how well they will fit into the group they have to work with (bad or good). These criteria are denoted, respectively, c_1, c_2, and c_3, and we set $X_{c_i} = \{-, +\}$, with $- \leq_{c_i} +$, for $i = 1, 2, 3$. Finally, some committee gives the candidates a global evaluation [B(ad), M(oderate) or G(ood)].

If we would run a classification tree algorithm on this problem, we would end up with one of the two trees depicted in Figure 1. If we choose tree T_1, for instance, it turns out that the best possible candidate, namely, someone with a lot of working experience, who is a fast learner and fits well in the group, in other words, a candidate with evaluations $(+, +, +)$, is evaluated as Moderate. However, another person who only has working experience, but needs much time to learn and will not get along with his colleagues, thus having evaluations $(+, -, -)$, ends up in the class labeled Good. This is in contradiction with the basic *principle of dominance preservation*, roughly stating that an object a with (partial) evaluations at most as good as the (partial) evaluations of an object b should get a global evaluation that also is at most as good.

DEFINITION 4. *The dominance relation*[d] $\lhd$ *on* Ω *with respect to* (*w.r.t.*) *a set of true criteria C is defined by*

[d]In the literature, the (strict) dominance relation usually is denoted by Δ. Because of the symmetric nature of the symbol Δ, we feel it does not clearly denote its meaning and we prefer the symbol $\lhd$.

Table I. Evaluations of candidates.

	c_1	c_2	c_3	f
a_1	−	−	+	B
a_2	−	+	−	M
a_3	+	−	+	G
a_4	+	+	−	M

$$a \lhd b \Leftrightarrow \begin{cases} (\forall\, c \in C)(c(a) \leq_c c(b)) \\ (\exists\, c \in C)(c(a) <_c c(b)) \end{cases}$$

for any $a, b \in \Omega$. It is said that a is dominated by b. If the second condition is not fulfilled, we say that a is weakly dominated by b and we write $a \trianglelefteq b$.

The principle of dominance preservation can now be formulated as

$$a \lhd b \Rightarrow f(a) \leq_d f(b) \tag{1}$$

If this principle is violated, we speak of reversed preference: there exists objects a, $b \in \Omega$ such that

$$a \lhd b \quad \text{and} \quad f(a) \not\leq_d f(b)$$

4. PARTIAL DOMINANCE

Example (Part 2). If we look at Table I, we see that there is no violation against the principle of dominance preservation. Nevertheless, uncareful induction of the ranking f may lead to (partial) reversed preferences. Let us have a closer look at how this can happen. If we analyze the possibilities for the first split, we notice that only the first option has some kind of monotone behavior as can be expected from a true criterion (Figure 2). More specifically, we could say that the first option preserves *partial* dominance in the sense that $c_1(a) <_{c_1} c_1(b) \Rightarrow f(a) \leq_d f(b)$. However, the split based on c_2 results in $c_2(a_3) <_{c_2} c_2(a_2) = c_2(a_4)$, and $f(a_3) =$

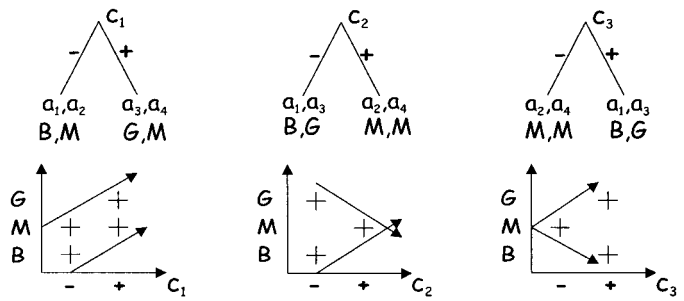

Figure 2. Three possible first splits.

$G \not\preceq_d f(a_2) = f(a_4) = M$. A similar counter-intuitive situation occurs for the split based on c_3.

Before introducing the notion of *partial dominance relation* as introduced in Refs. 8 and 9, we need to remark that in these works, a strong version of the principle of dominance preservation is used, namely,

$$a \trianglelefteq b \Rightarrow f(a) \preceq_d f(b)$$

Because it is based on the weaker version of the dominance relation in its antecedent, the principle becomes harder to satisfy; it demands that two objects a, $b \in \Omega$ with identical representation in the measurement space X should also have the same ranking, i.e., $f(a) = f(b)$. However, we would like to make a distinction between the fact that two vectors in X have different labels, a situation that we call doubt, and the fact that two objects exhibit reversed preference.

In Refs. 8 and 9 the (weak) partial dominance relation $\trianglelefteq_I$ on Ω w.r.t. a set of true criteria $C = \{c_i | i \in N\}$ and a subset $I \subseteq N$ is defined by

$$a \trianglelefteq_I b \Leftrightarrow (\forall\, c \in C_I)(c(a) \preceq_c c(b)) \tag{2}$$

for any a, $b \in \Omega$, where $C_I = \{c_i | i \in I\}$. Based on this definition, the principle of partial dominance preservation would become

$$a \trianglelefteq_I b \Rightarrow f(a) \preceq_d f(b)$$

It can be interpreted simply as the principle of dominance preservation making abstraction of the criteria that are not under consideration. For example, in choosing the first split in a tree-based approach, only the information of one criterion is taken into account, not the possible interactions between several criteria. However, as we continue expanding the tree, we soon realize that the principle of partial dominance preservation based on Equation 2 is no longer sufficient. We need to generalize this principle even further.

Our goal is to establish a comparison between two objects a and b based on whatever information is available about either of them. For instance, we know the partial evaluations of a for some subset $C_I \subseteq C$ and the partial evaluations of b for the criteria in $C_J \subseteq C$. But more general situations may occur; we might only know for an object a that for each $c \in C_I$ it holds that $c(a) \in V_c \subseteq X_c$, where V_c is not restricted to be a singleton as previously.

We can realize a comparison between two objects on the basis of such partial information by understanding more profoundly the idea underlying the principle of partial dominance preservation. In essence, it tries to establish a global comparison between two objects based on partial information. Because an objective global comparison can only be done based on the dominance relation, for which all information is required, we have to express partial dominance in terms of dominance that on its turn is expressed in terms of the product order $\preceq_X$ on X.

This leads us to the following definition (we first give an informal one, and then a formal one):

DEFINITION 5. *Let C be a set of true criteria and assume we only consider for each object in* Ω *the information available (in terms of partial evaluations). If a and b cannot be distinguished based on this available information, we say they are partially indifferent and write* $a -_p b$, *where the subscript "p" stands for "partial." In other words,* $a -_p b$ *is equivalent with "possibly* $(\forall\, c \in C)(c(a) = c(b))$*." The partial dominance relation* $\triangleleft_p$ *on* Ω *w.r.t. the set C and the information available is then defined by*

$$a \triangleleft_p b \Leftrightarrow \begin{cases} \text{possibly } a \triangleleft b \\ \text{impossibly } b \triangleleft a \\ \text{not } a -_p b \end{cases}$$

for any $a, b \in \Omega$. *The weak partial dominance relation* $\trianglelefteq_p$ *only fulfills the first two conditions.*

More formally, let π_1 *and* π_2 *be two subsets of X, i.e.,* π_1 *and* π_2 *correspond to sets of possible evaluations for objects from* Ω. *We define*

$$\pi_1 -_p \pi_2 \Leftrightarrow \pi_1 \cap \pi_2 \neq \varnothing \tag{3}$$

and

$$\pi_1 \triangleleft_p \pi_2 \Leftrightarrow \begin{cases} (\exists\, x_a \in \pi_1)(\exists\, x_b \in \pi_2)(x_a <_X x_b) \\ \neg(\exists\, x_a \in \pi_1)(\exists\, x_b \in \pi_2)(x_b <_X x_a) \\ \neg(\pi_1 \,{}_p \pi_2) \end{cases}$$

Assume now that we only have the information that the evaluation of object a belongs to π_1 *and the evaluation of object b belongs to* π_2, *then, we write* $a -_p b$ *(respectively,* $a \triangleleft_p b$*) if and only if* $\pi_1 -_p \pi_2$ *(respectively,* $\pi_1 \triangleleft_p \pi_2$*).*

In general, the previous definition tells us only how to compare two objects. One has to be careful in using these relations because neither $_p$ nor $\triangleleft_p$ are transitive. Let us show this by a simple example.

Example of the partial dominance relation. Assume we are working with two criteria with values in $X_{c_i} = \{0, 1, 2, 3\}$, where $0 <_{c_i} 1 <_{c_i} 2 <_{c_i} 3$. In this example, we use the following kind of shorthand writing: for $j \in X_{c_1}$, we set

$$(j, *) = \{a \in \Omega | c_1(a) = j,\ c_2(a) \in X_{c_2}\}$$

Now, let $a \in (1, 2)$, $b \in (2, *)$, $c \in (3, *)$, $d \in (3, 1)$, and $e \in (*, 2)$. We have that possibly $a \triangleleft b$ because $(1, 2) \triangleleft (2, 2)$ and $(2, 2)$ is a possible evaluation for b. On the other hand, it is impossible that $b \triangleleft a$ because $c_1(a) = 1 <_{c_1} c_1(b) = 2$. Clearly, $(1, 2) \cap (2, *) = \varnothing$, and hence we find $(1, 2) \triangleleft_p (2, *)$, implying $a \triangleleft_p b$. Likewise, we have $b \triangleleft_p c$, and $a \triangleleft_p c$. Yet, $\triangleleft_p$ is not transitive: $a \triangleleft_p b$ and $b \triangleleft_p d$, but not $a \triangleleft_p d$ because $c_1(a) <_{c_1} c_1(d)$ and $c_2(a) >_{c_2} c_2(d)$.

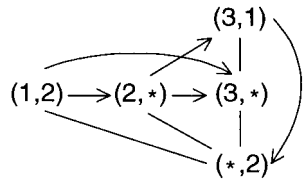

Figure 3. An arrow corresponds to $\lhd_p$, a line with $_p$.

The same is true for $-_p$ because $d -_p c$ and $c -_p e$, but $d \lhd_p e$. We have visualized all relations in Figure 3.

To show the use of $\trianglelefteq_p$, we mention that if $f \in (3, 0)$, then $f \trianglelefteq_p c$ [remember that $c \in (3, *)$], but not $f \lhd_p c$ because $f -_p c$. Remark, however, that $f \trianglelefteq_p c$ gives us more information than just $f -_p c$. In other words, $\trianglelefteq_p$ is not equivalent with $(\lhd_p \vee -_p)$. In fact, $\trianglelefteq_p$ and $-_p$ define two intersecting subsets of $X \times X$, where $\trianglelefteq_p \backslash -_p$ corresponds to $\lhd_p$ and $_{-p}\backslash \trianglelefteq_p$ is not necessarily empty.

This behavior is a consequence of the fact that $\lhd_p$ and $_{-p}$ try to approximate the relations $\lhd$ and $=$ based on incomplete information. If there is no information, all objects are treated the same, and distinction only starts when information becomes available. If all partial evaluations are precise, $\lhd_p$ and $-_p$ coincide with $\lhd$ and $=$. We also would like to mention that because a tree is a partition-based structure, the partial indifference relation $_{-p}$ will behave as an identity relation (on the leafs) in combination with $\lhd_p$, excluding most of the bizarre configurations mentioned in the previous example. Indeed, because the leaves form a partition, we find instead of Equation 3,

$$\pi_1 -_p \pi_2 \Leftrightarrow \pi_1 = \pi_2$$

for all leaves π_1 and π_2 of the tree. Moreover, if the splits are linear, we have that $\trianglelefteq_p$ corresponds to $(\lhd_p \vee -_p)$.

The principle of partial dominance preservation is a straightforward extension of Equation 1,

$$a \lhd_p b \Rrightarrow f(a) \preceq_d f(b)$$

This principle can be interpreted as follows: "if for two objects $a, b \in \Omega$, we know that based on the information we have (or take into account) about a and b, a might be dominated by b, then we consider a to be, at most, as good as b."

It can be shown that if the information available for objects a and b are the partial evaluations w.r.t. a set of criteria $C_I \subseteq C$ for a and the partial evaluations w.r.t. a set of criteria $C_J \subseteq C$ for b, then, if $C_I \cap C_J \neq \emptyset$,

$$a \trianglelefteq_p b \Leftrightarrow (\exists\, c \in C_I \cap C_J)(c(a) \preceq_c (b))$$

$$\Leftrightarrow a \trianglelefteq_{I \cap J} b$$

This shows that Equation 2 can be interpreted as using only information concerning a common subset of criteria ($C_I = C_J$) for establishing the partial dominance relation.

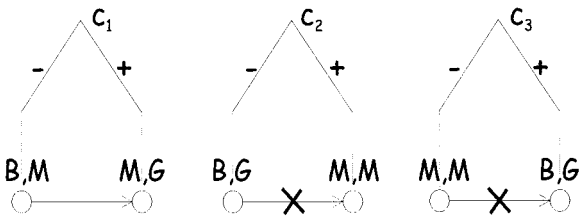

Figure 4. The splits based on c_1 and c_2 along with the partial dominance relation $\trianglelefteq_p$. The arrows corresponds to $\triangleleft_p$. We draw a cross over it if the principle of partial dominance preservation is violated.

Example (Part 3). We already explained why the first split should be based on c_1. With the definition of the partial dominance relation, it is now also possible to show graphically that the split based on c_1 is better than a split based on c_2 or c_3, as can be seen in Figure 4.

The possibilities for the second split are depicted in Figure 5. We clearly see that the split based on c_3 leads to a violation of the principle of partial dominance preservation: $a_2 \triangleleft_p a_1$ and $f(a_2) = \mathrm{M} \leq_d f(a_1) = \mathrm{B}$. The split based on criterion c_2 is in line with the principle and therefore is chosen. Finally, the last split is based on c_3 as can be seen in Figure 6.

In this way, we end up with the following rule base:

- If the candidate has little or no working experience and if, moreover, (s)he is a slow learner, then (s)he gets the global evaluation *bad.*
- If the candidate has little or no working experience but can compensate this a bit by being a fast learner, then (s)he is evaluated *moderate.*
- If the candidate has a lot of working experience, but doesn't fit well into the group, then (s)he is evaluated *moderate.*
- If the candidate has a lot of working experience combined with a good fit into the group, then (s)he is evaluated *good.*

This rule base is very natural, especially compared with the rule bases induced from Figure 1.

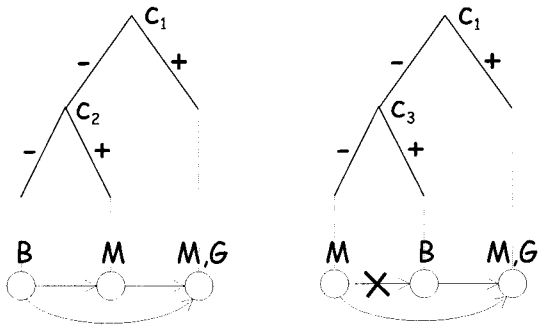

Figure 5. Two possibilities for the second split.

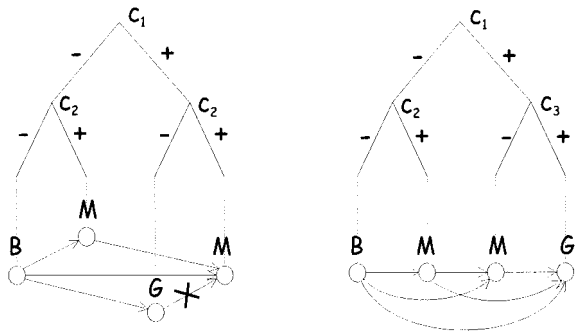

Figure 6. Two possibilities for the third split.

5. A REAL-WORLD EXAMPLE

The example elaborated in the previous sections is a purely academic one. Its simplicity hides several other difficulties inherent to the problem. First, the example has the characteristic that there is always just one split that does not violate the principle of partial dominance preservation. Of course, this is not a general property. In fact, quite often all splits will violate this principle, but some splits will do it more severely than others, causing more occurrences of (partial) reversed preference. This observation leads to a very rough measure γ for determining the best split: choose the split containing the least number of pairs of objects that violate the principle of partial dominance preservation. Although this heuristic is liable to many criticisms, it already seems to provide good results on small data sets. As is done frequently in machine learning, we could opt for a combined measure, e.g.,

$$\gamma(T) = \text{impurity}(T) + r \cdot \gamma'(T)$$

where some impurity measure is chosen and combined by a parameter $r > 0$ (expressing the relative importance of monotonicity) with a scoring measure γ' derived from γ, for example, in the same way the order ambiguity score of Ben-David (see Ref. 3) is constructed. However, we are reluctant to do so, because it is not clear whether such a linear combination has any actual meaning. Until more research has been done in this direction, we opt to leave the question of which measure to use unaddressed.

Second, the choice of the leaf that will be expanded can make a difference because the leaves are interconnected by the partial dominance relation $\trianglelefteq_p$. We propose to choose the leaf that is the least *pure*.[e] This choice can be defended by noticing that it helps in gaining a faster decrease of overall impurity. Remark that none of the existing "monotone" tree-based methods takes into account the fact that nonmonotonicity can be reduced more adequately by choosing the best node to

[e]Impurity can, e.g., be measured by the Shannon entropy or the Gini diversity index. See Ref. 14 for more details.

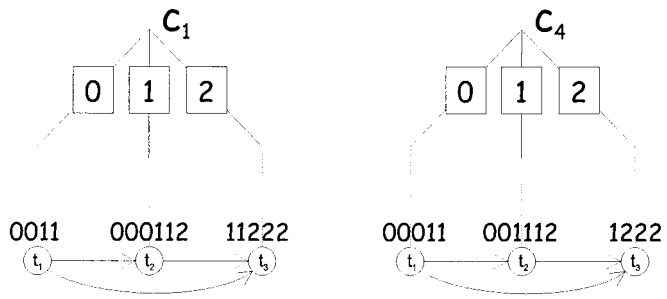

Figure 7. First split based on to c_1 and c_4.

split. This is because of the fact that they[f] only use a local splitting criterion as is usual in tree growing. The catch is that whereas the local optimum at a node coincides with the global optimum for the whole tree when considering classification trees, this is not necessarily true for ranking trees.

A third problem is that the final tree, and hence the resulting rule base, may contain reversed preferences, doubt (more than one evaluation is attached to a node) and even empty nodes. The solution to these problems is discussed theoretically in Ref. 18. In this section, we will use a real-world problem to guide us through the various problems and their solutions.

For demonstration purposes, we will use the data set published in Ref. 10, which is an adaption from a data set in Ref. 19. The data are shown in Table II, where 0 = low, 1 = moderate, and 2 = high. The aim is to predict the evaluation of countries on criterion d (*use of contraception*) based on the partial evaluations on the following criteria:

- Average years of education (c_1)
- Urbanization (c_2)
- Gross national product (GNP) per capita (c_3)
- Expenditure on family planning (c_4)

and to find those characteristics that are most valuable to predict d. Remark that this data set is not monotone, although nearly. In our approach, we keep all the available information; this is in contrast to the example in Ref. 4 where some information is deleted to ensure the monotonicity of the data set.

5.1. Growing the Ranking Tree

Considering the four possibilities for the first split, we find that the split based on c_1 leads to eight violations against the principle of partial dominance preservation, as can be seen in Figure 7; there are six violations between nodes t_1 and t_2,

[f]Except for the article of Ben-David,[2] which is rather an ad hoc solution to the problem of monotonicity, without bothering too much about the details.

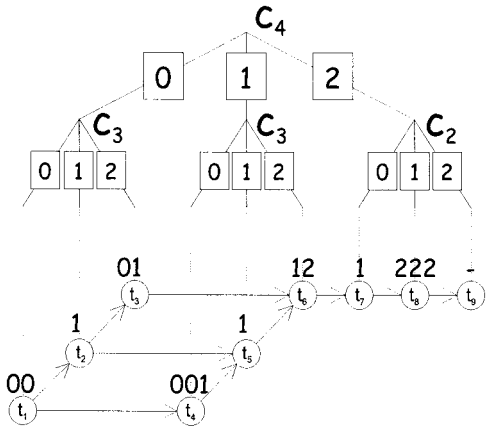

Figure 8. Contraception tree.

two violations between nodes t_2 and t_3, and no violations between nodes t_1 and t_3. The split based on c_2 causes 10 occurrences of reversed preference, c_3 gives the amount of 7 violations; and, finally, c_4 has a total of 5 violations. Therefore, we choose c_4 as the criterion to base the first split on.

Next, we search for the most impure node. In this example, we try to avoid overfitting in a naive fashion by not choosing nodes that contain less than four objects. In this way, node t_2 is identified as the most impure one. Even without any calculations, it is clear that this node is indeed the least accurate in labeling objects. Criteria c_1 and c_2 both leads to six violations against the principle of partial dominance preservation, and c_3 only lead to five violations. Criterion c_3 is chosen for this split. Now, t_1 has become the most impure node containing at least four objects. Splitting it according to c_1 results in seven violations, in five violations according to c_2, but the split based on c_3 only gives rise to two violations. Again, criterion c_3 is chosen. Last, node t_3 is chosen and this time criterion c_2 leads to the best results. The final tree is depicted[g] in Figure 8.

5.2. Building the Rule Base

Deriving a proper rule base from this tree still poses some problems that are connected with the assignment of a class label to each terminal node. First, it can be remarked that leaf t_9 does not contain any objects: there is no example of an object $a \in \Omega$ with $c_2(a) = c_4(a) = 2$. At first instance, this seems to imply that this node cannot be labeled. However, based on the principle of partial dominance preservation, we see that objects in this node should be labeled at least as high as objects from node t_8 that have partial evaluations $c_2 = 1$ and $c_4 = 2$. We conclude that node t_9 should be assigned to the evaluation 2 (high).

[g]Because the partial dominance relation is transitive in this particular case, the related graph is drawn as a transitive one in order to make the figure more transparent.

Second, some nodes (t_3, t_4, and t_6) bear more than one class label. Normally, some procedure is followed to choose one specific class label, e.g., the one that minimizes misclassification cost or, maybe even more adequate, the risk functional for ordinal regression derived by Herbrich.[13] However, in the context of ranking, information might get lost this way. Therefore, we opt to admit sets of class labels to be assigned. In Ref. 18 we argued that not all sets of class labels are meaningful and that only intervals are satisfactory. Indeed, an assignment of an object to, e.g., $\{0, 2\}$ is not meaningful in the context of ranking; if there is some hesitation between evaluating an object between low and high, one also should hesitate about assigning it to moderate. Hence, the set of labels used in the rule base must belong to the set $\mathscr{D}^{[2]} = \{[r, s] | (r, s) \in \mathscr{D}^2 \vee r \leq_d s\}$. This means that $\{0, 2\}$ should be replaced by the interval $[0, 2]$ in $(\mathscr{D}, \leq_d)$.

This leads to the following definition for the function $\hat{f}$ that assigns class labels to terminal nodes in the case of a ranking tree T:

$$\begin{aligned} \hat{f} : \tilde{T} &\rightarrow \mathscr{D}^{[2]} \\ t &\rightarrow \hat{f}(t) = [\hat{l}(t), \hat{r}(t)] \end{aligned} \tag{4}$$

where $\tilde{T}$ is the set of terminal nodes of the tree T, and

$$\hat{l}(x) = \inf\{f(a) \in \mathscr{D} | a \in \Omega \wedge a \in t\}$$

$$\hat{r}(x) = \sup\{f(a) \in \mathscr{D} | a \in \Omega \wedge a \in t\}$$

Next, because the final rule base should be an approximation of the ranking we were learning in the first place, a meaningful order $\leq_D$ needs to be defined on $\mathscr{D}^{[2]}$. Based on a long discussion about the semantics, we found in Ref. 18 that this order has to be defined as the well-known order that turns the set of intervals $\mathscr{D}^{[2]}$ into a lattice; for $D = [r_1, r_2]$, $D' = [s_1, s_2] \in \mathscr{D}^{[2]}$, we define

$$[r_1, r_2] \leq_D [s_1, s_2] \Leftrightarrow (r_1 \leq s_1) \wedge (r_2 \leq s_2)$$

In particular, when $r_1 = r_2 = r$ and $s_1 = s_2 = s$, we find $\{r\} \leq_D \{s\} \Leftrightarrow r \leq_d s$. In essence, this order means that if an object a is assigned to $[r_1, r_2]$ and an object b is assigned to $[s_1, s_2]$, then a might belong to the class with label r_1, which is worse than any of the classes object b might belong to, and, moreover, a can never belong to a class that is better than the best class b can belong to, so shifting from $[r_1, r_2]$ to $[s_1, s_2]$ means an improvement. It also is worth noting that $\leq_D$ does not imply risk aversion or risk proneness. The ranking $(f, \leq_d)$ used for expanding the tree is now approximated by $(\hat{f}, \leq_D)$.

We remark that the use of intervals and the foregoing order was already advocated in Ref. 16, but there, the intervals refer to the minimal and maximal monotone extensions of the (monotone) ranking and have no deeper inherent meaning. The fact that they used the same (and best known) order on the intervals therefore is merely a nice coincidence. In addition, it should be noted that if more a priori knowledge is available, e.g., risk aversion or risk proneness, other interval orders might become more appropriate (see Ref. 18). We could also use distribu-

tion functions for labeling, as long as stochastic dominance is guaranteed (see Ref. 18).

The rule base derived thus far might not be consistent as is the case in our example: because we have that $t_2 \lhd_p t_3$, the principle of partial dominance preservation dictates that we should have $\hat{f}(t_2) \leq_D \hat{f}(t_3)$, which is not the case in $(1 \not\leq_D [0, 1])$.

To eliminate reversed preference inside the rule base in the most minimal way and without discarding any information, we developed in Ref. 18 a function $\tilde{f}$ that results in a consistent rule base:

$$\tilde{f}(t) := [\min_{t' \in [t)} \hat{l}(t'), \max_{t' \in (t]} \hat{r}(t')]$$

where $[t) = \{t' \in \tilde{T} | t \trianglelefteq_p t'\}$ and $(t] = \{t' \in \tilde{T} | t' \trianglelefteq_p t\}$.

Remark that in spite of the seemingly close resemblance with the definitions of Potharst,[6] there is in fact a huge difference. First, Potharst is only interested in a much stricter version of these functions, defining them only on the original lattice $(\mathcal{D}^{[2]}, \leq_D)$. Moreover, because he only considers monotone data sets, the previous restriction implies he does not have to deal with any possible occurrences of reversed preference.

A second remark concerns the monotonicity of our solution. Any tree that is relabeled in this manner becomes monotone. Hence, the issue is not to construct a monotone tree (which is even not always possible in our setting), but to establish a tree expansion guideline that strives for monotonicity (in addition to purity). The principle of partial dominance preservation is such a guideline.

Applying the function $\tilde{f}$ to our example, we end up with the following assignments:

	$\{t_1\}$	$\{t_2, t_3, t_4\}$	$\{t_5\}$	$\{t_6, t_7\}$	$\{t_8\}$
$\tilde{f}$	0	$[0, 1]$	1	$[1, 2]$	2

or, restated as a rule base, where the "*don't care*" symbol $*$ denotes that the corresponding value is irrelevant:

IF				THEN
c_1	c_2	c_3	c_4	d
$*$	$*$	0	0	0
$*$	$*$	1	0	$[0,1]$
$*$	$*$	2	0	
$*$	$*$	0	1	
$*$	$*$	1	1	1
$*$	$*$	2	1	$[1,2]$
$*$	0	$*$	2	
$*$	1	$*$	2	2
$*$	2	$*$	2	

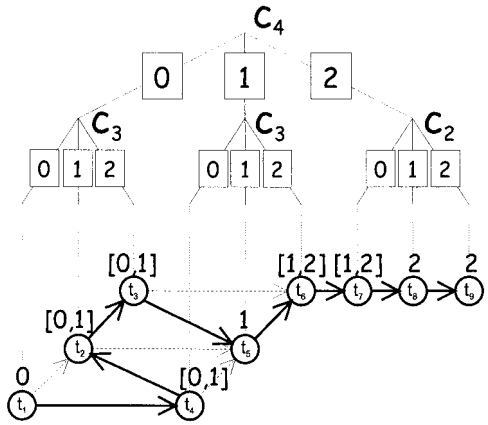

Figure 9. Visual rule base for "use of contraception."

5.3. Further Refinement of the Rule Base

The previous rule base can be synthesized into the following general statements:

- Expenditure on family planning (c_4) is the most important criterion.
- As long as c_4 is low or moderate, the GNP/capita (c_3) will influence the results.
- As soon as c_4 is high, the degree of urbanization (c_2) starts playing a distinctive role.

This conclusion still leaves open one major question: "How does the GNP/capita influence the results if the expenditure on family planning is low or moderate?" In other words, how do c_3 and c_4 interact if c_4 is low or moderate?

The solution to this problem lies in refining the class label assignments and their order even further. In essence, this can be done by taking into account the relative occurrence of each class label inside the terminal nodes. In this way, we can, e.g., easily recognize that a country belonging to t_4 is more likely to have a low use of contraception compared with a country belonging to t_3. Like this, we can see more clearly how the use of contraception increases in function of the considered criteria. This is shown in Figure 9, visually depicting the interaction between c_3 and c_4.

There are two striking elements in Figure 9. First, the interaction between c_3 and c_4 is not linear. Second, it suggests that a low GNP/capita (c_3) is a good indicator for a low use of contraception. This is confirmed by taking another look at Table II. This seems to contradict the fact that a fictitious country with evaluation (1, 1, 0, 2) is ranked as a country with a high use of contraception. However, this country is fictitious indeed because a high expenditure on family planning cannot be paired with a low GNP/capita in the real world (this can also be seen clearly in Table II; if c_4 is high, then c_3 is at least moderate).

As a final remark, we emphasize the fact that none of the existing methods will lend itself to a more refined analysis as we just did. The fact that we are able

Table II. Recoded contraception data.

Country	c_1	c_2	c_3	c_4	d
1. Lesotho	1	0	0	0	0
2. Kenya	0	0	0	0	0
3. Peru	1	1	2	0	0
4. Sri Lanka	1	1	0	1	0
5. Indonesia	0	0	0	1	0
6. Thailand	1	0	0	1	1
7. Colombia	1	2	1	1	1
8. Malaysia	0	1	2	1	1
9. Guyana	2	1	2	0	1
10. Jamaica	2	0	2	2	1
11. Jordan	0	2	1	0	1
12. Panama	2	2	2	1	2
13. Costa Rica	2	1	2	2	2
14. Fiji	1	1	2	2	2
15. Korea	2	1	1	2	2

c_1 = Average years of education; c_2 = urbanization; c_3 = GNP capita; c_4 = expenditure on family planning; d = Use of contraception; 0 = low; 1 = moderate; 2 = high.

to perform such an analysis is because of our graphical representation with a semantic interpretation. Note that this labeled partial dominance graph is our most interesting output, rather than the developed tree that is more a guiding tool for reading the graph.

6. CONCLUSION AND FURTHER RESEARCH

In the field of data mining, there exist many validated methods and algorithms for classification problems in which a classification has to be understood or learned on the basis of data. However, the same is not true for ranking problems.

In this study, we have addressed some important problems that classification algorithms face when they are confronted with a ranking problem, and we have shown, in particular, how these problems can be solved in tree-based approaches using the principle of partial dominance preservation. This has resulted in an easily interpretable and intuitive rule base that can even be presented graphically, unfolding instantly the existing complex interactions between the criteria.

An important question is how a meaningful measure can be built combining impurity and reversed preference. The fact that this is not an easy question can be seen in Ref. 10 where a measure in the framework of ranking is developed. Unfortunately, these results cannot be used in the construction of a tree.

References

1. Ben-David A. Automatic generation of symbolic multiattribute ordinal knowledge-based DSSs: Methodology and applications. Decis Sciences 1992;23:1357–1372.

2. Ben-David A, Sterling L, Pao YH. Learning and classification of monotonic ordinal concepts. Computat Intell 1989;5:45–49.
3. Ben-David A. Monotonicity maintenance in information-theoretic machine learning algorithms. Mach Learn 1995;19:29–43.
4. Makino K, Suda T, Ono H, Ibaraki T. Data analysis by positive decision trees. IEICE Trans Inf Syst 1999;E82-D:76–88. available at http://search.ieice.or.jp/1999/pdf/e82-d_1_76.pdf
5. Potharst R. Classification using decision trees and neural networks. PhD thesis, Erasmus University Rotterdam, 1999.
6. Potharst R, Bloch JC. Decision trees for ordinal classification. Intell Data Anal 2000;4: 97–112.
7. Breiman L. Using convex pseudo-data to increase prediction accuracy. Technical Report 513, Statistics Department University of California, Berkeley, available at http://oz.berkeley.edu/tech-reports/, 1998.
8. Greco S, Mattarazo B, Slowiński R. A new rough set approach to evaluation of bankruptcy risk. In: Zopounidis C, editor. Operational tool of the management of financial risk. Dordrecht: Kluwer; 1998. pp 121–136.
9. Greco S, Mattarazo B, Slowiński R. Rough set theory for multicriteria decision analysis. Eur J Oper Res 2000;129:1–47.
10. Düntsch I, Gediga G. Approximation quality for sorting rules. Computational Statistics and Data Analysis 2002;40:499–526.
11. Bioch JC, Popova V. Bankruptcy prediction with rough sets. Technical Report ERS-2001-11-LIS, Erasmus University Rotterdam, 2001.
12. Herbrich R, Graepel T, Obermayer K. Regression models for ordinal data: A machine learning approach. Technical Report TR 99-3, Technical University of Berlin, available at http://research.microsoft.com/users/rherb/techreports.htm, 1999.
13. Herbrich R, Graepel T, Obermayer K. Large margin rank boundaries for ordinal regression. In: Smola A, Bartlett P, Schölkopf B, Schuurmans D, editors. Advances in large margin classifiers. Cambridge, MA: MIT Press; 2000. pp 115–132.
14. Breiman L, Friedman J, Olshen R, Stone C. Classification and regression trees. New York: Chapman & Hall; 1984.
15. Vapnik V. Statistical learning theory. New York: John Wiley and Sons; 1998.
16. Roy B. Multicriteria methodology for decision aiding. Dordrecht: Kluwer Academic Publishers; 1996.
17. Bouyssou D. Building criteria: A prerequisite for MCDA. In: Bana e Costa C, editor. Readings in multiple criteria decision aid. Heidelberg: Springer-Verlag; 1990. pp 58–80.
18. Cao-Van K, De Baets B. Consistent representation of rankings. Submitted to European J. Oper. Res.
19. Cliff N. Predicting ordinal relations. Br J Math Stat Psychol 1994;47:127–150.