Classification Trees for Problems with Monotonicity Constraints

R. Potharst
Erasmus University Rotterdam
P.O. Box 1738
3000 DR Rotterdam
Netherlands
potharst@few.eur.nl

A. J. Feelders
Utrecht University
P.O. Box 80089
3508 TB Utrecht
Netherlands
ad@cs.uu.nl

ABSTRACT

For classification problems with ordinal attributes very often the class attribute should increase with each or some of the explaining attributes. These are called classification problems with monotonicity constraints. Classical decision tree algorithms such as CART or C4.5 generally do not produce monotone trees, even if the dataset is completely monotone. This paper surveys the methods that have so far been proposed for generating decision trees that satisfy monotonicity constraints. A distinction is made between methods that work only for monotone datasets and methods that work for monotone and non-monotone datasets alike.

Keywords

monotone, monotonicity constraint, classification, classification tree, decision tree, ordinal data

1. INTRODUCTION

Even though data mining is often applied to domains where little theory is available, in many cases it is either known that the target function satisfies certain constraints, or it is simply required that the model constructed satisfies those constraints.

One type of constraint that is available in many applications states that the dependent variable (or its expected value) should be a monotone function of the independent variables. Economic theory would state for example that people tend to buy less of a product if its price increases (ceteris paribus), so price elasticity of demand should be negative. The strength of this relationship and the precise functional form are however usually not dictated by economic theory. Other well-known examples are labor wages as a function of age and education (see e.g. [11]) or so-called hedonic price models where the price of a consumer good depends on a bundle of characteristics for which a valuation exists [9].

Another class of problems where monotonicity constraints often apply are so-called selection problems, such as the selection of applicants for a job or a loan on the basis of their characteristics. As an example, consider a selection procedure for applicants to a job based on the outcomes of a series of academic and/or psychological tests. If each of the test outcomes x_i is scored from low (bad performance) to high (good performance) and the classes are taken to be 0 = not selected and 1 = selected, then it would be very natural to demand the selection rule to be monotone. In fact, the requirement of monotonicity would be equivalent to excluding all situations in which applicant A scores better or at least as good on all tests as applicant B, whereas B gets selected and A does not.

Because the monotonicity constraint is quite common in practice, many data analysis techniques have been adapted to be able to handle such constraints.

Isotonic regression, for example, deals with regression problems with monotonicity constraints. The traditional method used in isotonic regression is the *pool-adjacent violators algorithm* [15]. This method however only works in the one-dimensional case. A versatile non-parametric method is given in [11].

Monotonicity constraints have also been investigated in the neural network literature. In [16] the monotonicity of the neural network is guaranteed by enforcing constraints on the weights during the training process. Daniels and Kamp [8] present a class of neural networks that are monotone by construction. This class is obtained by considering multi-layer neural networks with non-negative weights.

Various methods have also been proposed for classification problems with monotonicity constraints, such as decision lists [4], logical analysis of data [5], rough sets [6] and instance-based learning [3; 1].

Classification or decision trees are among the most popular algorithms for classification problems in data mining and machine learning. Therefore we consider in this paper methods to build monotone classification trees.

In Section 2 we define monotone classification and other important concepts that are used throughout the paper. We also provide a motivating example concerning applicants for a bank loan, that is used to illustrate many of the algorithms presented.

The paper then divides into algorithms that work on monotone datasets (Section 3) and algorithms that also work on non-monotone data sets (Section 4).

In Section 3.2 we present an algorithm that forces the construction of a monotone tree by adding, if required, the corner elements of a node with an appropriate class label to the dataset.

In Section 4 we present two algorithms that work on non-monotone data. The first is due to Ben-David [2], and

adapts the well-known entropy splitting criterion by including a measure for the non-monotonicity of the tree that results after the split. In Section 4.2 we present a straightforward generate-and-test approach that constructs many different trees by resampling the training data, and selects a monotone tree.
Finally, in Section 5 we end with a discussion, and some ideas for further research.

2. MONOTONE CLASSIFICATION

Let $\mathcal{X}$ be a partially ordered set of instances, called the *instance space*, and let $\mathcal{C}$ be a finite linearly ordered set of *classes*. The order relations of $\mathcal{X}$ and $\mathcal{C}$ will both be denoted by $\leq$. An *allocation rule* is a function

$$f : \mathcal{X} \to \mathcal{C}$$

which assigns a class from $\mathcal{C}$ to every instance in the instance space $\mathcal{X}$. A *classification problem* is the problem of finding a class labeling f that satisfies certain constraints, to be specified in the problem description. One possible constraint is that the labeling f be monotone: a *monotone* allocation rule is a function $f : \mathcal{X} \to \mathcal{C}$ for which

$$\mathbf{x} \leq \mathbf{x}' \Rightarrow f(\mathbf{x}) \leq f(\mathbf{x}') \tag{1}$$

for all instances $\mathbf{x}, \mathbf{x}' \in \mathcal{X}$. In this paper, $\mathcal{X}$ will always be a *feature space* $\mathcal{X} = \mathcal{X}_1 \times \mathcal{X}_2 \times \ldots \times \mathcal{X}_p$ consisting of vectors $\mathbf{x} = (x_1, x_2, \ldots, x_p)$ of values on p features or attributes. Here we assume that each feature takes values x_i in a linearly ordered set $\mathcal{X}_i$. The partial ordering $\leq$ on $\mathcal{X}$ will be the ordering induced by the order relations of its coordinates $\mathcal{X}_i$: $\mathbf{x} = (x_1, x_2, \ldots, x_p) \leq \mathbf{x}' = (x'_1, x'_2, \ldots, x'_p)$ if and only if $x_i \leq x'_i$ for all i. It is easy to see that a classification rule on a feature space is monotone if and only if it is non-decreasing in each of its features, when the remaining features are held fixed.
A very common classification problem occurs, when the allocation rule should be induced from an available dataset or set of examples: for a finite number of instances a corresponding class is given; an allocation rule should be constructed that 'fits' these data. Formally, a *dataset* is a series $(\mathbf{x}_1, c_1), (\mathbf{x}_2, c_2), \ldots, (\mathbf{x}_n, c_n)$ of n examples $(\mathbf{x}_i, c_i)$ where each $\mathbf{x}_i$ is an element of the instance space $\mathcal{X}$ and c_i is a class label from $\mathcal{C}$. The presence of noise may lead to inconsistencies in the dataset that might disturb the faultless operation of our algorithms. We call a dataset *consistent* if for all i, j we have $\mathbf{x}_i = \mathbf{x}_j \Rightarrow c_i = c_j$. That is, each instance in the dataset has a unique associated class. For such a dataset it makes sense to speak of the class $\lambda(\mathbf{x})$ associated with an instance $\mathbf{x}$. Another important distinction we make in this paper is between monotone and non-monotone datasets. In fact, the methods of Section 3 work only for monotone datasets whereas those of Section 4 can be used also for non-monotone datasets. We call a dataset *monotone* if for all i, j we have $\mathbf{x}_i \leq \mathbf{x}_j \Rightarrow c_i \leq c_j$. It is easy to see that a monotone dataset is necessarily consistent. In fact, if $\mathbf{x}_i = \mathbf{x}_j$ then we have $\mathbf{x}_i \leq \mathbf{x}_j$ and $\mathbf{x}_j \leq \mathbf{x}_i$, so $c_i \leq c_j$ and $c_j \leq c_i$, and consequently, $c_i = c_j$. This discussion leads to the following formal definitions.

DEFINITION 1. A *consistent dataset* $\mathcal{D}$ is a pair (D, λ) where $D \subset \mathcal{X}$ is a finite subset of the instance space $\mathcal{X}$ and $\lambda : D \to \mathcal{C}$ is a class labeling of the elements of D. The pairs $(\mathbf{x}, \lambda(\mathbf{x}))$ with $\mathbf{x} \in D$ will be called the *examples* of the dataset.

Note that the class labeling λ of a consistent dataset $\mathcal{D} = (D, \lambda)$ is *not* an allocation rule: it is only defined on D, a subset of $\mathcal{X}$, while an allocation rule must be defined on all elements of the instance space $\mathcal{X}$. In fact, a classification problem for a consistent dataset consists of finding an allocation rule f that is an extension of the class labeling λ of the dataset to the whole instance space $\mathcal{X}$.

DEFINITION 2. A *monotone dataset* is a consistent dataset $\mathcal{D} = (D, \lambda)$ for which the implication (1) holds for all $\mathbf{x}, \mathbf{x}' \in D$ with f replaced by λ.

We will now give an example of a monotone classification problem. Suppose a bank wants to base its loan policy on a number of features of its clients, for instance on income, education level and criminal record. If a client is granted a loan, it can be one in three classes: low, intermediate and high. So, together with the 'no loan' option, we have four classes. Suppose further that the bank wants to base its loan policy on a number of credit worthiness decisions in the past. These past decisions are given in Table 1:

client	*income*	*education*	*crim.record*	*loan*
cl1	low	low	fair	no
cl2	low	low	excellent	low
cl3	average	intermediate	excellent	intermediate
cl4	high	low	excellent	high
cl5	high	intermediate	excellent	high

Table 1: The bank loan dataset

A client with features at least as high as those of another client may expect to get at least as high a loan as the other client. So, finding a loan policy compatible with past decisions amounts to solving a monotone classification problem with the dataset of Table 1.
In order to save space we will often map the values of the attributes of a dataset to a set of numbers. For instance, Table 1 could be written as

X_1	X_2	X_3	C
0	0	1	0
0	0	2	1
1	1	2	2
2	0	2	3
2	1	2	3

when we use the mapping low $\to$ 0, average $\to$ 1, high $\to$ 2 for feature $X_1 =$ *income*, etc. More often, we will write concisely

001	0
002	1
112	2
202	3
212	3

for the above dataset.
Finally, we will establish some notation to be used throughout this paper:

- The minimal and maximal elements of $\mathcal{C}$ will be denoted by $c_{\min}$ and $c_{\max}$ respectively.
- $[a, b]$ denotes the interval $\{\mathbf{x} \in \mathcal{X} : a \leq \mathbf{x} \leq b\}$, where both a and b are instance vectors from $\mathcal{X}$.
- $(a, b]$ denotes the interval $\{\mathbf{x} \in \mathcal{X} : a < \mathbf{x} \leq b\}$, where both a and b are instance vectors from $\mathcal{X}$.
- For all $\mathbf{x} \in \mathcal{X}$, we define the *upset* generated by $\mathbf{x}$ as

  $$\uparrow\mathbf{x} = \{\mathbf{y} \in \mathcal{X} : \mathbf{y} \geq \mathbf{x}\}$$

  and, if D is a subset of $\mathcal{X}$ the upset generated by D is defined as

  $$\uparrow D = \bigcup_{\mathbf{x} \in D} \uparrow\mathbf{x}.$$

- Similarly, for $\mathbf{x} \in \mathcal{X}$, we define the *downset* generated by $\mathbf{x}$ as

  $$\downarrow\mathbf{x} = \{\mathbf{y} \in \mathcal{X} : \mathbf{y} \leq \mathbf{x}\}$$

  and the downset generated by a subset D of $\mathcal{X}$ is defined as

  $$\downarrow D = \bigcup_{\mathbf{x} \in D} \downarrow\mathbf{x}.$$

2.1 Monotone Extensions of Datasets

As noted above the problem of finding a solution to a monotone classification problem amounts to finding a monotone extension f of the class labeling λ of a dataset $\mathcal{D} = (D, \lambda)$. Formally, a function $f : \mathcal{X} \to \mathcal{C}$ is an *extension* of $\lambda : D \to \mathcal{C}$, if the restriction of f to D i.e. $f|D = \lambda$. Or, if $f(\mathbf{x}) = \lambda(\mathbf{x})$ for all $\mathbf{x} \in D$. If $\mathcal{D} = (D, \lambda)$ is monotone, we denote the collection of all monotone extensions of λ with $M(\mathcal{D})$. Note that $M(\mathcal{D})$ is partially ordered by the order relation $f \leq f'$ iff $f(\mathbf{x}) \leq f'(\mathbf{x})$ for all $\mathbf{x} \in \mathcal{X}$. We will now define two special elements of this collection.

Definition 3. If $\mathcal{D} = (D, \lambda)$ is a monotone dataset, we define $\lambda^{\mathcal{D}}_{\min} : \mathcal{X} \to \mathcal{C}$, and $\lambda^{\mathcal{D}}_{\max} : \mathcal{X} \to \mathcal{C}$, as follows: for all $\mathbf{x} \in \mathcal{X}$

$$\lambda^{\mathcal{D}}_{\min}(\mathbf{x}) = \begin{cases} \max\{\lambda(\mathbf{y}) : \mathbf{y} \in D \cap \downarrow\mathbf{x}\} & \text{if } \mathbf{x} \in \uparrow D \\ c_{\min} & \text{otherwise} \end{cases}$$

and

$$\lambda^{\mathcal{D}}_{\max}(\mathbf{x}) = \begin{cases} \min\{\lambda(\mathbf{y}) : \mathbf{y} \in D \cap \uparrow\mathbf{x}\} & \text{if } \mathbf{x} \in \downarrow D \\ c_{\max} & \text{otherwise.} \end{cases}$$

We will now show[1] that the functions $\lambda^{\mathcal{D}}_{\min}$ and $\lambda^{\mathcal{D}}_{\max}$, as defined, are the minimal resp. maximal elements of $M(\mathcal{D})$.

Lemma 1. *If $\mathcal{D} = (D, \lambda)$ is a monotone dataset, for the functions $\lambda^{\mathcal{D}}_{\min}$ and $\lambda^{\mathcal{D}}_{\max}$ the following statements hold:*

(i) $\lambda^{\mathcal{D}}_{\min}, \lambda^{\mathcal{D}}_{\max} \in M(\mathcal{D})$

(ii) $M(\mathcal{D}) = \{f : \lambda^{\mathcal{D}}_{\min} \leq f \leq \lambda^{\mathcal{D}}_{\max}$ *and* f *monotone*$\}$.

Theoretically, we now have at least two solutions for a monotone classification problem with dataset $\mathcal{D} = (D, \lambda)$: the minimal and maximal extension of λ. These two allocation rules we will call the *minimal rule* and the *maximal rule* respectively. In addition we have for every point $\mathbf{x}$ in the instance space bounds that any rule f must satisfy:

$$\lambda^{\mathcal{D}}_{\min}(\mathbf{x}) \leq f(\mathbf{x}) \leq \lambda^{\mathcal{D}}_{\max}(\mathbf{x}).$$

Any monotone allocation rule that satisfies these bounds will be another solution to our problem.
In Section 3 we will require the representation of our allocation rule to have a specific form, viz. the form of a classification tree or decision tree.

3. METHODS FOR MONOTONE DATA

Classification or decision trees have long been used for classification problems. Well-known introductions to this field can be found in [7] and [14]. In this paper we will only consider so-called *univariate* decision trees: at each split the decision to which of the disjoint subsets an element belongs, is made using the information from one feature or attribute only. Within this class of univariate decision trees, we will only consider so-called *binary* trees. For such trees, at each node a split is made using a test of the form

$$X_i \leq c \ (\text{or} X_i < c)$$

for some $c \in \mathcal{X}_i, 1 \leq i \leq n$. Thus, for a binary tree, in each node[2] the associated set $T \subset \mathcal{X}$ is split into the two subsets $T_\ell = \{\mathbf{x} \in T : x_i \leq c\}$ and $T_r = \{\mathbf{x} \in T : x_i > c\}$. An example of a univariate binary decision tree is the following:

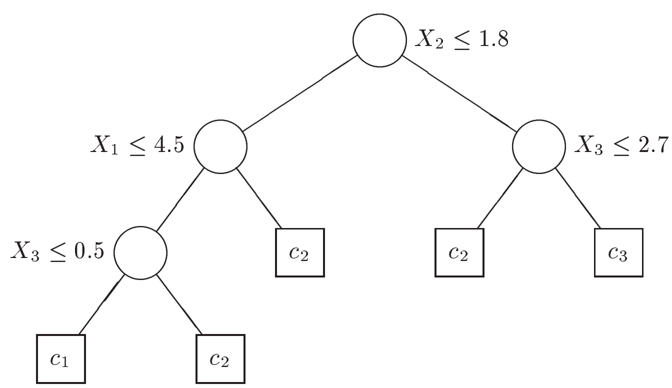

Figure 1: Univariate Binary Decision Tree: Example

This tree splits the instance space $\mathcal{X} = \mathbb{R}^3$ into the five regions

$$\begin{aligned} T_1 = &\quad \{\mathbf{x} \in \mathbb{R}^3 : x_1 \leq 4.5, x_2 \leq 1.8, x_3 \leq 0.5\} \\ T_2 = &\quad \{\mathbf{x} \in \mathbb{R}^3 : x_1 \leq 4.5, x_2 \leq 1.8, x_3 > 0.5\} \\ T_3 = &\quad \{\mathbf{x} \in \mathbb{R}^3 : x_1 > 4.5, x_2 \leq 1.8\} \\ T_4 = &\quad \{\mathbf{x} \in \mathbb{R}^3 : x_2 > 1.8, x_3 \leq 2.7\} \\ T_5 = &\quad \{\mathbf{x} \in \mathbb{R}^3 : x_2 > 1.8, x_3 > 2.7\} \end{aligned}$$

the first and the last of which are classified as c_1 and c_3 respectively, and the remaining regions as c_2. The allocation rule that is induced by a decision tree $\mathcal{T}$ will be denoted by $f_{\mathcal{T}}$.

[1] The proofs of all lemmas in this paper can be found in [12].

[2] By slight abuse of language in the sequel we will make no distinction between a node or leaf and its associated subset.

Lemma 2. *If $\mathcal{X}$ is an instance space with continuous features and $\mathcal{T}$ is a univariate binary decision tree on $\mathcal{X}$, then if $T \subset \mathcal{X}$ is the subset associated with an arbitrary node or leaf of $\mathcal{T}$,*

$$T = \{\mathbf{x} \in \mathcal{X} : a < \mathbf{x} \leq b\} = (a, b] \qquad (2)$$

for some $a, b \in \overline{\mathcal{X}}$ with $a \leq b$.

Here we use the expression $\overline{\mathcal{X}}$ instead of $\mathcal{X}$, because in some cases $\mathcal{X}$ would have to be extended with infinity-elements in order to have a representation of form (2) for each node or leaf.
If $\mathcal{X}$ is an instance space with discrete features, then any subset T associated with a univariate binary decision tree $\mathcal{T}$ on $\mathcal{X}$ will satisfy

$$T = \{\mathbf{x} \in \mathcal{X} : a \leq \mathbf{x} \leq b\} = [a, b] \qquad (3)$$

for some $a, b \in \mathcal{X}$, with $a \leq b$. As an abbreviation we will use the notation $T = [a, b]$ for a set of this form. Below we will call $\min(T) = a$ the *minimal element*[3] and $\max(T) = b$ the *maximal element* of T. Together, we call these the *corner* elements of the node T.

3.1 Testing the Monotonicity of a Decision Tree

In this subsection we describe an efficient algorithm for testing whether a given decision tree $\mathcal{T}$ is monotone or not. A naive way to test the monotonicity of a decision tree $\mathcal{T}$ would be to check all pairs of instances $\mathbf{x}, \mathbf{x}' \in \mathcal{X}$, determine $f_{\mathcal{T}}(\mathbf{x})$ and $f_{\mathcal{T}}(\mathbf{x}')$ by throwing them through the tree and check whether we find a non-monotonicity like $\mathbf{x} \leq \mathbf{x}'$ and at the same time $f_{\mathcal{T}}(\mathbf{x}) > f_{\mathcal{T}}(\mathbf{x}')$. Of course, this method would be very time consuming and, in the continuous case, even sheer impossible. Fortunately, there is a straightforward manner to test the monotonicity using the maximal and minimal elements of the leaves of the decision tree:

for all pairs of leaves T, T':
  if $\Big(f_{\mathcal{T}}(T) > f_{\mathcal{T}}(T')$ **and** $\min(T) < \max(T')\Big)$ **or**
    $\Big(f_{\mathcal{T}}(T) < f_{\mathcal{T}}(T')$ **and** $\max(T) > \min(T')\Big)$
  then stop: $\mathcal{T}$ not monotone

It is easy to check that a decision tree is passed through the above algorithm without stopping, if and only if the tree is monotone.

3.2 The Direct Method

In this subsection we will describe the algorithm proposed in [12] for the induction of a monotone binary decision tree from a monotone dataset. The algorithm has been tested extensively on artificial and real world data, see [13] for an application to a bankruptcy problem. We will first describe the algorithm for the case of a discrete feature space. At the end of the section we will indicate what changes are needed to run this algorithm in the continuous case.
Since this is a survey article on monotone decision trees, it should be mentioned that in addition to this direct method of constructing a monotone decision tree, there is also an indirect method: first, construct a so-called quasi-monotone tree, and subsequently, if needed, repair the quasi-monotone tree if it still contains non-monotonicities. This indirect method was proposed in [10] for two-class problems and generalized to multi-class problems in [12]. Although the concept of quasi-monotonicity is theoretically satisfying, the indirect method is found to be lacking efficiency compared to the direct method that is treated in this subsection.
An algorithm for the induction of a decision tree $\mathcal{T}$ from a dataset $\mathcal{D}$ contains the following ingredients:

- a *splitting rule* $\mathcal{S}$: defines the way to generate a split in each node,
- a *stopping rule* $\mathcal{H}$: determines when to stop splitting and form a leaf,
- a *labeling rule* $\mathcal{L}$: assigns a class label to a leaf when it is decided to create one.

If $\mathcal{S}, \mathcal{H}$ and $\mathcal{L}$ have been specified, then an *induction algorithm* according to these rules can be recursively described as in Figure 2.

tree($\mathcal{X}, \mathcal{D}_0$):
  split($\mathcal{X}, \mathcal{D}_0$)

split(T, **var** $\mathcal{D}$):
  $\mathcal{D}$:= update($\mathcal{D}, T$);
  if $\mathcal{H}(T, \mathcal{D})$ **then**
    assign class label $\mathcal{L}(T, \mathcal{D})$ to leaf T
  else
    begin
      $(T_\ell, T_r) := \mathcal{S}(T, \mathcal{D})$;
      split $(T_\ell, \mathcal{D})$;
      split $(T_r, \mathcal{D})$
    end

Figure 2: Monotone Tree Induction Algorithm

In this algorithm outline there is one aspect that we have not mentioned yet: the *update rule*. In the algorithm we use, we shall allow the dataset to be updated at various moments during tree generation. During this process of updating we will incorporate in the dataset knowledge that is needed to guarantee the monotonicity of the resulting tree.
Note, that $\mathcal{D}$ must be passed to the split procedure as a *variable* parameter, since $\mathcal{D}$ is updated during execution of the procedure.
In addition to the update rule, we need to specify a splitting rule, a stopping rule and a labeling rule. Together these are then plugged into the algorithm of Figure 2 to give a complete description of the algorithm under consideration.
We start with describing the update rule. When this rule fires, the dataset $\mathcal{D} = (D, \lambda)$ will be updated: at most two elements will be added to the dataset, each time the update rule fires. As soon as a node T is accessed, either the minimal element of T or the maximal element, or both will be added to $\mathcal{D}$, provided with a well-chosen class labeling. If both these corner elements of T already belong to D, nothing changes. The complete update rule is given in Figure 3.
When a minimal element of node T is added to the dataset, it gets the highest possible class label. In contrast, a maximal element that is added to the dataset will receive the lowest possible class label. The reason for this choice has to do with the desire to produce a small tree. It speeds up the course towards homogeneous leaves.

[3] In the continuous case this definition implies $\min(T) \notin T$, but that does not lead to any complications.

```
update (var D, T):
    a := min(T);
    b := max(T);
    if a ∉ D then
        begin
            λ(a) := λ^D_max(a);
            D := D ∪ {a}
        end;
    if b ∉ D then
        begin
            λ(b) := λ^D_min(b);
            D := D ∪ {b}
        end;
    return D = (D, λ)
```

Figure 3: The Standard Update Rule

The splitting rule $\mathcal{S}(T, \mathcal{D})$ must be such that at each node the associated subset T is split into two nonempty subsets

$$\mathcal{S}(T,\mathcal{D}) = (T_\ell, T_r) \quad \begin{array}{l} \text{with } T_\ell = \{\mathbf{x} \in T : x_i \leq c\} \\ \text{and } T_r = \{\mathbf{x} \in T : x_i > c\} \end{array} \tag{4}$$

for some $i \in \{1, \ldots, p\}$, and some $c \in \mathcal{X}_i$. Furthermore, the splitting rule must satisfy the following requirement: i and c must be chosen such that

$$\exists \mathbf{x}, \mathbf{x}' \in D \cap T \text{ with } \lambda(\mathbf{x}) \neq \lambda(\mathbf{x}'), \mathbf{x} \in T_\ell \text{ and } \mathbf{x}' \in T_r. \tag{5}$$

Next, we consider the stopping rule $\mathcal{H}(T, \mathcal{D})$. As a result of the actions of the update rule, both the minimal element $\min(T)$ and the maximal element $\max(T)$ of T belong to D. Now, as a stopping rule we will use:

$$\mathcal{H}(T,\mathcal{D}) = \begin{cases} \textbf{true} & \text{if } \lambda(\min(T)) = \lambda(\max(T)), \\ \textbf{false} & \text{otherwise.} \end{cases} \tag{6}$$

Finally, the labeling rule $\mathcal{L}(T, \mathcal{D})$ will be simply:

$$\mathcal{L}(T,\mathcal{D}) = \lambda(\min(T)) = \lambda(\max(T)). \tag{7}$$

For the proof that this algorithm works we will need two lemmas. The first of these lemmas tells us that if we add an instance to a dataset while giving it a class label that is in between the lower and upper bounds that are given by the dataset as it is now, the dataset remains monotone. The second lemma tells us that if the minimal and maximal element of a node both have the same class label, then we can make this node into a leaf with that class label.

Lemma 3. *Let $\mathcal{D} = (D, \lambda)$ be a monotone dataset with $D \subset \mathcal{X}$ and $\lambda : D \to \mathcal{C}$. Let $\mathbf{x}^+$ be an arbitrary instance vector with $\mathbf{x}^+ \notin D$, and let $c \in \mathcal{C}$ be such that*

$$\lambda^{\mathcal{D}}_{\min}(\mathbf{x}^+) \leq c \leq \lambda^{\mathcal{D}}_{\max}(\mathbf{x}^+).$$

If $\mathcal{D}^+ = (D^+, \lambda^+)$ is defined as follows:

$$\begin{cases} D^+ = D \cup \{\mathbf{x}^+\} \\ \lambda^+(\mathbf{x}) = \begin{cases} \lambda(\mathbf{x}) & \text{for } \mathbf{x} \in D \\ c & \text{for } \mathbf{x} = \mathbf{x}^+ \end{cases} \end{cases}$$

then the following assertions are true:

(i) *$\mathcal{D}^+$ is a monotone dataset,*

(ii) $\lambda^{\mathcal{D}}_{\min} \leq \lambda^{\mathcal{D}^+}_{\min} \leq \lambda^{\mathcal{D}^+}_{\max} \leq \lambda^{\mathcal{D}}_{\max}$

(iii) $M(\mathcal{D}^+) \subset M(\mathcal{D})$.

(iv) $Q(\mathcal{D}^+) \subset Q(\mathcal{D})$.

Lemma 4. *If $\mathcal{D} = (D, \lambda)$ is a monotone dataset and $a, b \in D$, such that $a \leq b$ and $\lambda(a) = \lambda(b) = c \in \mathcal{C}$, then for all monotone allocation rules $f \in M(\mathcal{D})$ we have for all $\mathbf{x} \in T = \{\mathbf{x} \in \mathcal{X} : a \leq \mathbf{x} \leq b\}$*

$$f(\mathbf{x}) = c.$$

Now we can formulate and prove the main theorem of this section.

Theorem 1. *Let $\mathcal{X}$ be a finite instance space with discrete features and let $\mathcal{D} = (D, \lambda)$ be a monotone dataset on $\mathcal{X}$. If the functions $\mathcal{S}, \mathcal{H}, \mathcal{L}$ satisfy the requirements (4),(5),(6) and (7), then the algorithm of Figure 2 together with the update rule of Figure 3 will generate a monotone decision tree $\mathcal{T}$ with $f_{\mathcal{T}} \in M(\mathcal{D})$.*

Proof: The update rule of the algorithm generates a finite sequence of datasets $\mathcal{D}_1, \mathcal{D}_2, \ldots, \mathcal{D}_k$, with $\mathcal{D}_i = (D_i, \lambda_i)$, $D_i \in \mathcal{X}$, $\lambda_i : D_i \to \mathcal{C}$, $1 \leq i \leq k$, such that, according to Lemma 3, each $\mathcal{D}_i$ is monotone, $D \subset D_1 \subset D_2 \subset \ldots \subset D_k$,

$$\lambda^{\mathcal{D}}_{\min} \leq \lambda^{\mathcal{D}_1}_{\min} \leq \ldots \leq \lambda^{\mathcal{D}_k}_{\min} \leq \lambda^{\mathcal{D}_k}_{\max} \leq \ldots \leq \lambda^{\mathcal{D}_1}_{\max} \leq \lambda^{\mathcal{D}}_{\max},$$

and

$$M(\mathcal{D}_k) \subset \ldots \subset M(\mathcal{D}_1) \subset M(\mathcal{D}).$$

The update rule guarantees, that the minimal and maximal element of each node, where the stopping rule fires, are members of the dataset. For such a node, Lemma 4 asserts there is only one labeling possible. For the last dataset $\mathcal{D}_k$ we must have: all minimal and maximal elements of all leaves are members of $\mathcal{D}_k$, so $M(\mathcal{D}_k)$ will consist of just one member: $f_{\mathcal{T}}$. The process must be finite since we have a finite instance space $\mathcal{X}$, and each D_i must be a subset of $\mathcal{X}$. □

Note, that this theorem actually proves a whole class of algorithms to be correct: the requirements set by the theorem to the splitting rule are quite general. Nothing is said in the requirements about how to select the attribute X_i and how to calculate the cut-off point c for a test of the form $t = \{X_i \leq c\}$. Obvious candidates for attribute-selection and cut-off point calculation are the well-known impurity measures like entropy, Gini or the twoing rule, see [7].

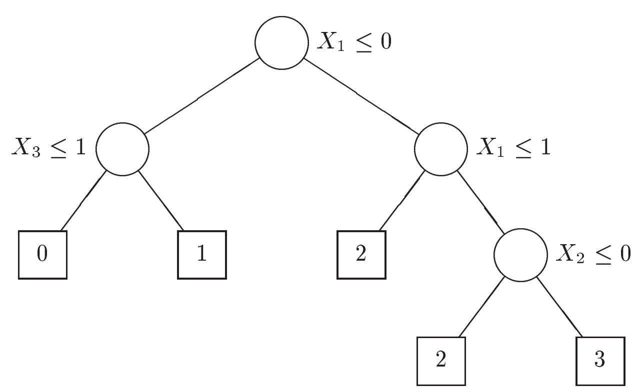

Figure 4: Monotone Decision Tree for the Bank Loan Dataset

As an illustration of the operation of the presented algorithm we will use it to generate a monotone decision tree for the dataset of Table 1. As an impurity criterion we will use entropy, see [14]. Starting in the root, we have $T = \mathcal{X}$, so $a = 000$ and $b = 222$. Now, $\lambda^{\mathcal{D}}_{\max}(000) = 0$ and $\lambda^{\mathcal{D}}_{\min}(222) = 3$, so the elements 000:0 and 222:3 are added to the dataset, which then consists of 7 examples. Next, six possible splits are considered: $X_1 \leq 0, X_1 \leq 1, X_2 \leq 0, X_2 \leq 1, X_3 \leq 0$ and $X_3 \leq 1$. For each of these possible splits we calculate the decrease in entropy as follows. For the test $X_1 \leq 0$, the space $\mathcal{X} = [000, 222]$ is split into the subset $T_\ell = [000, 022]$ and $T_r = [100, 222]$. Since T_ℓ contains three data elements and T_r contains the remaining four, the average entropy of the split is $\frac{3}{7} \times 0.92 + \frac{4}{7} \times 1 = 0.97$. Thus, the decrease in entropy for this split is $1.92 - 0.97 = 0.95$. When calculated for all six splits, the split $X_1 \leq 0$ gives the largest decrease in entropy, so it is used as the first split in the tree. Proceeding with the left node $T = [000, 022]$ we start by calculating $\lambda^{\mathcal{D}}_{\min}(022) = 1$ and adding the element 022:1 to the dataset $\mathcal{D}$, which will then have eight elements. We then consider the four possible splits $X_2 \leq 0, X_2 \leq 1, X_3 \leq 0$ and $X_3 \leq 1$, of which the last one gives the largest decrease in entropy, and leads to the nodes $T_\ell = [000, 021]$ and $T_r = [002, 022]$. Since $\lambda^{\mathcal{D}}_{\min}(021) = 0 = \lambda(000)$, T_ℓ is made into a leaf with class 0. Proceeding in this manner we end up with the decision tree of Figure 4 which is easily checked to be monotone.

3.3 A variation

A useful variation of the above algorithm is the following. We change the update rule to

update (**var** $\mathcal{D}, T$):
    if T is homogeneous **then**
        begin
            $a := \min(T)$;
            $b := \max(T)$;
            if $a \notin D$ **then**
                begin
                    $\lambda(a) := \lambda^{\mathcal{D}}_{\max}(a)$;
                    $D := D \cup \{a\}$
                end;
            if $b \notin D$ **then**
                begin
                    $\lambda(b) := \lambda^{\mathcal{D}}_{\min}(b)$;
                    $D := D \cup \{b\}$
                end
        end

Figure 5: Update Rule: a variation

thus, only adding the minimal and maximal elements of a node T to the dataset if the node is *homogeneous*, i.e. if

$$\forall \mathbf{x}, \mathbf{y} \in D \cap T : \lambda(\mathbf{x}) = \lambda(\mathbf{y}).$$

The splitting rule, stopping rule and labeling rule remain the same. With these changes the theorem remains true as can be easily seen. However, whereas with the standard algorithm from the beginning one works at 'monotonizing' the tree, this algorithm starts adding corner elements only when it has found a homogeneous node. For instance, if one uses maximal decrease of entropy as a measure of the performance of a test-split $t = \{X_i \leq c\}$, this algorithm is equal to Quinlan's C4.5-algorithm, until one hits upon a homogeneous node; from then on our algorithm starts adding the corner elements $\min(T)$ and $\max(T)$ to the dataset, enlarging the tree somewhat, but making it monotone. We call this process *cornering*. Thus, the algorithm of Figure 5 can be seen as a method that first builds a traditional (non-monotone) tree with a method such as ID3, C4.5 or CART, and next makes it monotone by adding corner elements to the dataset. This observation yields also the possible use of this variant: if one has an arbitrary (non-monotone) tree for a monotone classification problem, it can be 'repaired' i.e. made monotone by adding corner elements to the leaves and growing some more branches where necessary.

As an example of the use of this remark, suppose we have the following monotone dataset $\mathcal{D}$:

000	0
001	1
100	0
110	1

Suppose further, that someone hands us the following decision tree for classifying the above dataset:

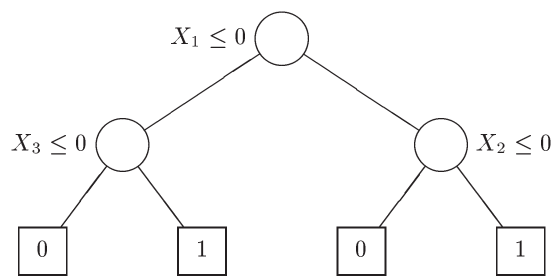

Figure 6: Non-monotone Decision Tree

This tree indeed classifies $\mathcal{D}$ correctly, but although $\mathcal{D}$ is monotone, the tree is not. In fact, it classifies data element 001 as belonging to class 1 and 101 as 0. Clearly, this is against monotonicity rule (1). To correct the above tree, we apply the algorithm of Figure 5 to it. We add the maximal element of the third leaf 101 to the dataset with the value $\lambda^{\mathcal{D}}_{\min}(101) = 1$. The leaf is subsequently split and the resulting tree is easily found to be monotone, see Figure 7. Of course, if we would have grown a tree directly with the above dataset $\mathcal{D}$ with the standard algorithm we would have ended up with a smaller tree, which is equally correct and monotone, see Figure 8. Nevertheless, it helps to know that we can make an arbitrary tree monotone by splitting up some of the leaves and adding a few more branches.

The main algorithm of this section further suggests a new impurity measure to be used as an attribute selection criterion. First note, that for each $T = \{\mathbf{x} \in \mathcal{X} : a \leq \mathbf{x} \leq b\}$ with $T \cap D \neq \emptyset$ we have

$$\lambda^{\mathcal{D}}_{\max}(a) \leq \lambda^{\mathcal{D}}_{\min}(b).$$

This can be seen as follows: let $\mathbf{x}_0$ be an element of $T \cap D$, then

$$\lambda^{\mathcal{D}}_{\max}(a) \leq \lambda(\mathbf{x}_0) \leq \lambda^{\mathcal{D}}_{\min}(b).$$

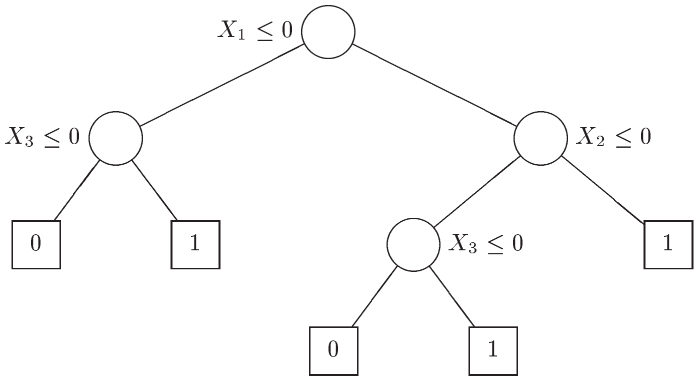

Figure 7: The above tree, but repaired

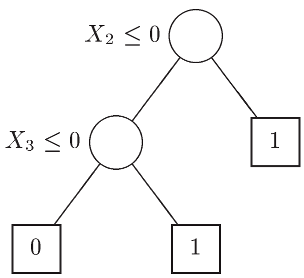

Figure 8: Monotone Tree produced by the Standard Algorithm

We now define the *variation* of the dataset on T as

$$\text{var }(T) = |[\lambda^{\mathcal{D}}_{\max}(a), \lambda^{\mathcal{D}}_{\min}(b)]| - 1,$$

the number of different class labels that are possible within node T minus one. It is clear that $\text{var}(T) = 0$ iff $\lambda^{\mathcal{D}}_{\max}(a) = \lambda^{\mathcal{D}}_{\min}(b)$. Clearly, this measure can be used as an impurity measure, and the decrease in variation can be taken as an attribute selection criterion. However, experiments have shown that it is inferior to entropy or Gini: trees grown with this impurity measure tend to be somewhat larger than those grown with entropy or the Gini-index.

3.4 Changes Needed for Continuous Attributes

Here we will sum up the changes that need to be made to the described algorithms in case one or more of the attributes is continuous. For simplicity of notation we will assume that all attributes $X_i, 1 \leq i \leq p$, are continuous on a finite or infinite subinterval $\mathcal{X}_i$ of $\mathbb{R}$. If in practice, some of the attributes are discrete while others are continuous, the reader can easily adapt the described procedures to that situation. Thus, we assume that we have an infinite instance space $\mathcal{X} = \mathcal{X}_1 \times \ldots \times \mathcal{X}_p$, with $\mathcal{X}_i$ a subinterval of $\mathbb{R}$, the set of real numbers. However, the dataset $\mathcal{D} = (D, \lambda)$ will always be finite. In particular, let us assume that attribute X_i has values

$$x_i^{(1)} < x_i^{(2)} < \ldots < x_i^{(k_i)}$$

in the dataset D, where k_i is the number of different values that attribute X_i has in the dataset D. Of course, $k_i \leq |\mathcal{D}|$. In fact, with probability one we have $k_i = |\mathcal{D}|$, but, because of rounding off, in practice $k_i < |\mathcal{D}|$ will often occur. Now, we define

$$\mathcal{X}_i^{\mathcal{D}} = \{x_i^{(1)}, \ldots, x_i^{(k_i)}\}$$

and

$$\mathcal{X}^{\mathcal{D}} = \mathcal{X}_1^{\mathcal{D}} \times \mathcal{X}_2^{\mathcal{D}} \times \ldots \times \mathcal{X}_p^{\mathcal{D}}.$$

Thus, $\mathcal{X}^{\mathcal{D}}$ is a finite space which includes all instances in D, and which is discrete. So we have mapped the classification problem with infinite instance space $\mathcal{X}$ onto a classification problem with finite space $\mathcal{X}^{\mathcal{D}}$. Using the methods of this section we can generate a decision tree for the classification problem on $\mathcal{X}^{\mathcal{D}}$. The final step then will be to translate this decision tree on $\mathcal{X}^{\mathcal{D}}$ to a decision tree on $\mathcal{X}$.
Let $\mathcal{T}$ be a binary monotone decision tree on $\mathcal{X}^{\mathcal{D}}$, generated by one of the methods of this section using dataset $\mathcal{D}$. Each test of this tree will have either the form

$$X_i \leq x_i^{(j)} \tag{8}$$

for some j with $1 < j \leq k_i$, for some $i \in \{1, \ldots, p\}$. With a test of the form (8) $j = k_i$ is impossible since in that case one of the splitted sets would be empty.
Now, we replace each test of the form (8) by

$$X_i \leq \frac{x_i^{(j)} + x_i^{(j+1)}}{2}.$$

These changes will give us a binary decision tree on $\mathcal{X}$ that classifies the dataset $\mathcal{D}$ correctly.
As an example, let us assume we have a dataset with one continuous attribute X_1, while all other attributes are discrete. Let us further assume that X_1 has values

0.51 0.98 1.43 2.87 3.11

in the dataset. With these values, seen as discrete values, a decision tree is built which happens to have two nodes in which X_1 plays a role: in one node we have a test $X_1 \leq 0.98$ and in the other node we have $X_1 \leq 2.87$. Both tests are subsequently replaced by $X_1 \leq (0.98+1.43)/2$ or $X_1 \leq 1.205$ and $X_1 \leq (2.87 + 3.11)/2$ or $X_1 \leq 2.99$ respectively. This is similar to applying a continuity correction when approximating a discrete distribution by a continuous distribution in statistics.
As a final remark, note that in practice it is usually advisable to discretize continuous attributes, since working with too many values per attribute leads to prohibitive computing times.

4. METHODS FOR NON-MONOTONE DATA

The algorithms discussed so far work for monotone datasets. Even if the true underlying relation is monotone, the observed data may, as a consequence of noise, not be. Furthermore, sometimes we simply require that the allocation rule be monotone, even if we believe that the underlying relation is not. In that case the task is to find a monotone model with good predictive performance.
In this section we look at two approaches that can handle non-monotone and inconsistent datasets.

4.1 The Weighted Sum Method

Ben-David [2], proposes a tree induction algorithm that is similar to well-known algorithms such as C4.5 and CART.

The important difference with those algorithms is that the splitting rule includes a measure of the degree of monotonicity of the tree in addition to the usual impurity measure.

To this end a $k \times k$ symmetric non-monotonicity matrix M is defined, where k equals the number of leaves of the tree constructed so far. The m_{ij} element of M equals 1 if leaf T_i is non-monotone with respect to leaf T_j and 0 otherwise. Clearly, the diagonal elements of M are 0. A non-monotonicity index I is defined as follows

$$I = \frac{W}{k^2 - k},$$

where W denotes the sum of M's entries, and $k^2 - k$ is the maximum possible value of W for any tree with k leaves [2]. Note however that this maximum can only be achieved if there are at least k distinct classes.

Based on this non-monotonicity index the order-ambiguity-score of a decision tree is defined as follows

$$A = \begin{cases} 0 & \text{if } I = 0 \\ -(\log_2 I)^{-1} & \text{otherwise} \end{cases}$$

Finally the splitting rule is redefined to include the order-ambiguity-score

$$S = E + \rho A,$$

where S denotes the total-ambiguity-score to be minimized, E is the well-known entropy measure, and ρ is a parameter that expresses the importance of monotonicity relative to inductive accuracy. The quality of each split is determined by computing its total-ambiguity-score, where A is the order-ambiguity-score of the tree that results from the split.

Note that W is a rather crude measure of the degree of non-monotonicity of a tree, since each non-monotone leaf pair has equal weight. A possible improvement would be to weight the different leaves according to their probability of occurrence. The matrix M' could now be defined as follows. The m'_{ij} element of M' equals $p(T_i) \times p(T_j)$ if leaf T_i is non-monotone with respect to leaf T_j and 0 otherwise, where $p(T_i)$ denotes the proportion of cases in leaf T_i. The non-monotonicity index becomes

$$I' = \frac{W'}{(k^2 - k)/k^2} = \frac{W'}{1 - 1/k},$$

where W' is again the sum of the entries of M', and the maximum is attained when all possible leaves are non-monotone with respect to each other and occur with equal probability $1/k$. W' is an estimate of the probability that if we draw two points at random from the feature space, these points turn out to lie in two leaves that are non-monotone with respect to each other. Note that $p(T_i) \times p(T_j)$ is an upper bound for the degree of non-monotonicity between node T_i and T_j because not all elements of T_i and T_j have to be non-monotone with respect to each other.

The most straightforward way to measure the degree of non-monotonicity of a tree would be to use it to label all data, and simply count the number of non-monotone pairs created by the labeling. This is however computationally rather demanding since this should be performed for the collection of trees that results from applying each possible split.

4.2 A Generate-and-Test Approach

The use of a measure of monotonicity in determining the best split, as discussed in the previous section, has certain drawbacks. Monotonicity is a *global* property, i.e. it involves a relation between different leaf nodes of a tree. If the degree of monotonicity is measured for each possible split during tree construction, the *order* in which nodes are expanded becomes important. For example, a depth-first search strategy will generally lead to a different tree then a breadth-first search. Also, and perhaps more importantly, a non-monotone tree may become monotone after additional splits.

In view of these drawbacks, we consider an alternative approach in this section. Rather than *enforcing* monotonicity during tree construction, we generate many different trees and *check* if they are monotonic. The collection of trees may be obtained by drawing bootstrap samples from the training data, or making different random partitions of the data in a training and test set. This approach allows the use of a standard tree algorithm except that the minimum and maximum elements of the nodes have to be recorded during tree construction, in order to be able to check whether the final tree is monotone. This approach has the additional advantage that one can estimate to what extent the assumption of monotonicity is correct, by looking at the proportion of monotone trees versus non-monotone trees obtained.

The tree algorithm used is in many respects similar to the CART program as described in [7]. The program makes binary splits and uses the Gini-index as splitting criterion. Furthermore it uses cost-complexity pruning [7] to generate a nested sequence of trees from which the best one is selected on the basis of test set performance. During tree construction, the algorithm records the minimum and maximum element for each node. These are used to check the whether a tree is monotone. In Figure 9 we give pseudo-code for the tree construction algorithm with recording of the corner elements of each node.

The function *leaf* determines whether a node should be turned into a leaf. This is the case when the node is homogeneous, all examples in the node have identical attribute values, or the node contains too few examples to be split any further. A class label is assigned to the leaf, by default based on the majority rule.

The minimum and maximum element of root node T_0 are set to $-\infty$ and ∞ respectively. The updating of corner elements proceeds as follows. The minimum of T_ℓ is identical to that of T, and the same goes for the maximum of T_r. For the maximum of T_ℓ and the minimum of T_r, x_{j^*} (the split attribute) is set to c^* (the split value) and for all other attributes they are the same as the maximum and minimum of T respectively.

Determining the non-monotone pairs of leaf nodes is straightforward: take any pair (T, T') with $f_{\mathcal{T}}(T) > f_{\mathcal{T}}(T')$ and check if $\min(T) < \max(T')$. If this is the case, then add (T, T') to the list of non-monotone leaf-pairs.

In the next section we illustrate this algorithm by applying it to an economic dataset concerning house prices.

4.3 Application to House Pricing

In this section we illustrate the resampling approach described in the previous section. We discuss its application to the prediction of the price-category of a house in the city of

```
growtree(𝒳,𝒟 : training sample):
    T_0 := 𝒳
    for i ∈ {1,...,p}
        min_i(T_0) := −∞
        max_i(T_0) := +∞
    split(T_0,min(T_0), max(T_0), 𝒟)

split(T, min(T), max(T), 𝒟):
    if leaf(T, 𝒟) then
      assign class label ℒ(T, 𝒟) to T
    else
      𝒮 := allsplits(T, 𝒟)
      (j*, c*) := arg max_{(j,c)∈𝒮} quality(j, c, T, 𝒟)
      T_ℓ := {x ∈ T : x_{j*} < c*}
      T_r := {x ∈ T : x_{j*} ≥ c*}
      for i ∈ {1,...,p}
          min_i(T_ℓ) := min_i(T)
          max_i(T_r) := max_i(T)
          if i = j*
              max_i(T_ℓ) := c*
              min_i(T_r) := c*
          else
              max_i(T_ℓ) := max_i(T)
              min_i(T_r) := min_i(T)
      split (T_ℓ, min(T_ℓ), max(T_ℓ), 𝒟)
      split (T_r, min(T_r), max(T_r), 𝒟)
```

Figure 9: Tree Induction Algorithm with recording of node corners

Den Bosch (a medium sized Dutch city with approximately 120,000 inhabitants).

The attributes $x_1, x_2, \ldots, x_p$ are characteristics of the house. They have been selected on the basis of interviews with experts of local house brokers, and advertisements offering real estate in local magazines. The monotonicity constraint makes sense for this application, since the better the characteristics of a house, the higher the asking price. The most important attributes are listed in Table 2.

It is a relatively small data set with only 119 observations used for illustrative purposes only. Of all 7021 distinct pairs of observations, 2217 are comparable, and 78 are non-monotonic. For the purpose of this study we have discretized the dependent variable (asking price) into the classes *below median* (euro 157,955) and *above median*. After this discretization of the dependent variable 9 pairs of observations are non-monotonic.

In order to determine the effect of application of the monotonicity constraint we repeated the following experiment 100 times. The dataset was randomly partitioned (within classes) into a training set (60 observations) and test set (59 observations). The training set was used to construct a sequence of trees using cost-complexity pruning. From this sequence the best tree was selected on the basis of error rate on the test set (in case of a tie, the smallest tree was chosen). Finally, it was checked whether the tree was monotone and if not, the upperbound W' for the degree of monotonicity (as defined in Section 4.1) was computed.

Out of the 100 trees thus constructed, 57 turned out to be monotone and 43 not. The average misclassification rate of the monotone trees was 14.93% against 14.94% for the non-monotone trees. Thus, the predicitive accuracy was virtually identical.

Symbol	Definition
DISTR	type of district
	four categories ranked from bad to good
SURF	total area including garden
RM	number of bedrooms
TYPE	1. apartment
	2. row house
	3. corner house
	4. semidetached house
	5. detached house
	6. villa
VOL	volume of the house
STOR	number of storeys
GARD	type of garden
	four categories ranked from bad to good
GARG	1. no garage
	2. normal garage
	3. large garage

Table 2: Definition of attributes for house pricing example

Figure 10 depicts one of the 43 non-montonic trees, and figure 11 one of the 57 monotone trees obtained in the experiment. Class label 0 corresponds to prices below the median and label 1 to prices above median. In figure 10, the number of a leaf is given directly below it. It is easily verified that the leaf-pairs (3,4), (3,7), (5,7) and (6,7) are non-monotone. The degree of non-monotonicity W' (see Section 4.1) of this tree is only about 1%. The tree in figure 11 is monotone and has only 3 leaf nodes. The estimated error of the non-monotone tree shown is 15.3%, and the estimated error of the monotone tree 13.6%.

The average degree of non-monotonicity W' of the non-monotone trees was about 1.6%, which is quite low, the more if we take into consideration that this is an upper bound. Another interesting comparison is between the average sizes of the trees. On average, the monotone trees had about 3.19 leaf nodes, against 6.95 for the non-monotone trees. Thus, the monotone trees are considerably smaller and therefore easier to understand. The variability around the mean number of leaf nodes can be used as a measure of the stability of the trees generated. For the monotone trees, the variance of the number of leaf nodes was 0.91 against 5.05 for the non-monotone trees. Clearly then the monotone trees are more stable upon repeated sampling than there non-monotone counterparts.

5. DISCUSSION

Monotonicity is a common type of constraint on models in data mining. Furthermore, monotonicity may be an important requirement for explaining and justifying model outcomes. We have investigated the use of monotonicity constraints in classification tree algorithms.

We have presented algorithms that work on monotone data only, as well as algorithms that work on both monotone and non-monotone data. The former could be made more widely applicable by developing sensible methods to make a non-monotone data set monotone by making as few adjustments

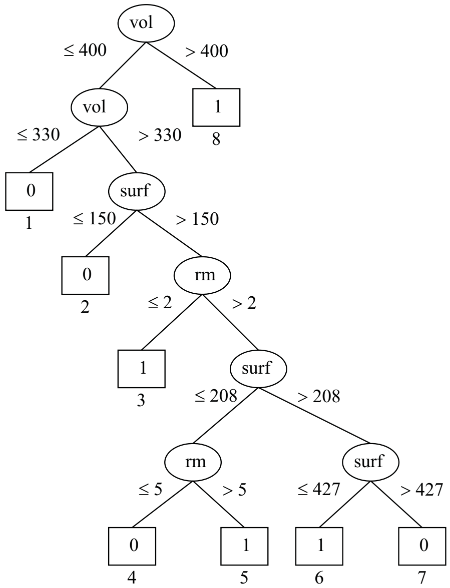

Figure 10: Example of a non-monotone tree

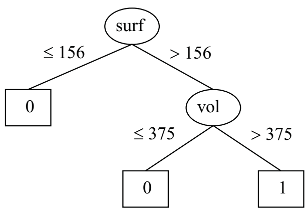

Figure 11: Example of a monotone tree

to the data as possible.

For non-monotone data we have presented two algorithms, the weighted-sum method and a generate-and-test algorithm. In preliminary experiments with the generate-and-test algorithm on house pricing data, we have found that the predictive performance of monotone trees was comparable to the performance of the non-monotone trees. However, the monotone trees were much simpler and therefore more insightful and easier to explain. Furthermore, the monotone trees proved to be more stable upon repeated sampling. This provides interesting prospects for applications where monotonicity is an absolute requirement, such as in many selection decision models.

An interesting, as yet unexplored, approach for non-monotone data would be to use a pruning method that prunes towards monotone subtrees of the initially grown tree. One could create a nested sequence of monotone subtrees of the initial tree, and select from this sequence the tree with the best predictive accuracy on a test set. Another interesting extension of the work surveyed in this paper is to consider multivariate classification trees, where each split may be based on more than one attribute.

6. REFERENCES

[1] A. Ben-David. Automatic generation of symbolic multi-attribute ordinal knowledge-based DSS's: methodology and applications. *Decision Sciences*, 23:1357–1372, 1992.

[2] A. Ben-David. Monotonicity maintenance in information-theoretic machine learning algorithms. *Machine Learning*, 19:29–43, 1995.

[3] A. Ben-David, L. Sterling, and Y. Pao. Learning and classification of monotonic ordinal concepts. *Computational Intelligence*, 5:45–49, 1989.

[4] J. Bioch. Dualization, decision lists and identification of monotone discrete functions. *Annals of Mathematics and Artificial Intelligence*, 24:69–91, 1998.

[5] J. Bioch and T. Ibaraki. Complexity of identification and dualization of positive Boolean functions. *Information and Computation*, 123:50–63, 1995.

[6] J. Bioch and V. Popova. Rough sets and ordinal classification. In A. S. H. Arimura, S. Jain, editor, *Algorithmic Learning Theory*, Lecture Notes in Artificial Intelligence 1968, pages 291–305. Springer, 2000.

[7] L. Breiman, J. Friedman, R. Olshen, and C. Stone. *Classification and Regression Trees (CART)*. Wadsworth, 1984.

[8] H. Daniels and B. Kamp. Application of MLP networks to bond rating and house pricing. *Neural Computation and Applications*, 8:226–234, 1999.

[9] O. Harrison and D. Rubinfeld. Hedonic prices and the demand for clean air. *Journal of Environmental Economics and Management*, 53:81–102, 1978.

[10] K. Makino, T. Susa, K. Yano, and T. Ibaraki. Data analysis by positive decision trees. In Y. Kambayashi et al., editors, *Proceedings of the International Symposium on Cooperative Database Systems for Advanced Applications (CODAS)*, pages 257–264, Kyoto, Japan, December 1996. World Scientific.

[11] H. Mukarjee and S. Stern. Feasible nonparametric esimation of multiargument monotone functions. *Journal of the American Statistical Association*, 89(425):77–80, 1994.

[12] R. Potharst. *Classification using Decision Trees and Neural Nets*. PhD thesis, Erasmus University Rotterdam, 1999.

[13] R. Potharst and J. Bioch. Decision trees for ordinal classification. *Intelligent Data Analysis*, 4(2):97–112, 2000.

[14] J. Quinlan. *C4.5 Programs for Machine Learning*. Morgan Kaufmann, 1993.

[15] T. Robertson, F. Wright, and R. Dykstra. *Order Restricted Statistical Inference*. Wiley, 1988.

[16] S. Wang. A neural network method of density estimation for univariate unimodal data. *Neural Computation & Applications*, 2:160–167, 1994.