Possibilistic Induction in Decision-Tree Learning

Eyke Hüllermeier

Department of Mathematics and Computer Science
University of Marburg, Germany
eyke@mathematik.uni-marburg.de

Abstract. We propose a generalization of Ockham's razor, a widely applied principle of inductive inference. This generalization intends to capture the aspect of uncertainty involved in inductive reasoning. To this end, Ockham's razor is formalized within the framework of possibility theory: It is not simply used for identifying a single, apparently optimal model, but rather for concluding on the possibility of various candidate models. The possibilistic version of Ockham's razor is applied to (lazy) decision tree learning.

1 Introduction

Inductive reasoning – by its very nature – is inseparably connected with *uncertainty* [4]. To begin with, the data presented to learning algorithms is imprecise, incomplete or noisy most of the time, a problem that can badly mislead a learning procedure. But even if observations are perfect, the generalization beyond that data is still afflicted with uncertainty. For example, observed data can generally be explained by more than one candidate theory, which means that one can never be sure of the truth of a particular model. In fact, the insight that inductive inference can never produce ultimate truth can be traced back at least as far as Francis Bacon's epistemology. In his Novum Organum[1], Bacon advocates a gradualist conception of inductive enquiry and proposes to set up *degrees of certainty*. Thus, from experience one may at best conclude that a theory is *likely* to be true – not, however, that it is true with full certainty.

In machine learning and mathematical statistics, uncertainty is often handled by means of probabilistic methods. In Bayesian approaches, for example, the data-generating process is modeled by means of a probability distribution which depends on the true model. Given the data S, a (posterior) probability (density) can thus be assigned to each model $M \in \mathcal{M}$, where $\mathcal{M}$ is the class of candidate models. The specification of a probability distribution, μ, over that class of models allows one to take the uncertainty related to the learning (prediction) task into account. For example, rather than making a single prediction $y_0 = M^*(x_0)$ on the basis of a particular model M^* (and a given query x_0), one can derive a probability $\mathsf{Pr}(y) = \mu(\{M \in \mathcal{M} \mid M(x_0) = y\})$ for each potential outcome y.

Probabilistic approaches are not always applicable, however, and they do not capture every kind of uncertainty relevant to machine learning. Particularly, this

[1] Published in 1620.

T. Elomaa et al. (Eds.): ECML, LNAI 2430, pp. 173–184, 2002.

appears to be true for the uncertainty or, say, unreliability connected to heuristic principles of inductive inference such as Ockham's razor. Such principles usually suggest one particular model $M^* \in \mathcal{M}$, thereby disregarding the aspect of uncertainty. Our aim in this paper is to alleviate this drawback by means of a *possibilistic* approach to inductive inference. More specifically, we shall propose a formalization of Ockham's razor within the framework of possibility theory. In its generalized version, Ockham's razor specifies the possibility of alternative models rather than selecting one particular model.

Section 2 recalls some basic principles of decision tree learning. In Section 3, the possibilistic version of Ockham's razor is introduced. The application of this generalized principle to classical decision tree learning and to a lazy variant thereof are discussed, respectively, in Sections 4 and 5. Finally, Section 6 presents some experimental results.

2 Decision Tree Learning

We proceed from the common framework for learning from examples: $\mathcal{X}$ denotes the instance space, where an instance corresponds to the description x of an object in attribute–value form. That is, each object x is characterized through attribute values $\alpha_\imath(x) \in A_\imath$, $1 \leq \imath \leq k$, where $A_\imath = \mathrm{dom}(\alpha_\imath)$ is the (finite) domain of the $\imath$-th attribute $\alpha_\imath$; the set of all attributes is denoted $\mathcal{A}$. $\mathcal{L} = \{\lambda_1, \ldots, \lambda_m\}$ is a set of labels, and $\langle x, \lambda_x \rangle$ is called a labeled instance or an example. S denotes a sample that consists of n labeled instances $\langle x_\imath, \lambda_{x_\imath} \rangle$, $1 \leq \imath \leq n$. Finally, a new instance (query) $x_0 \in \mathcal{X}$ is given, whose label λ_{x_0} is to be estimated.

The basic principle underlying most decision tree learners, well-known examples of which include the ID3 algorithm [12] and its successor C4.5 [13] as well as the CART system [2], is that of partitioning the set of given examples, S, in a recursive manner. Each inner node η of a decision tree τ defines a partition of a subset $\mathcal{S}_\eta \subset \mathcal{S}$ of examples assigned to that node. This is done by classifying elements $x \in \mathcal{S}_\eta$ according to the value of a specific attribute α. The attribute is selected according to a measure of effectiveness in classifying the examples, thereby supporting the overall objective of constructing a small tree. A widely applied "goodness of split" measure is the *information gain*, $G(S, \alpha)$, which is defined as the expected reduction in entropy (impurity) which results from partitioning S according to α:

$$G(S, \alpha) \doteq \mathrm{ent}(S) - \sum_{u \in \mathrm{dom}(\alpha)} \frac{|S_u|}{|S|} \, \mathrm{ent}(S_u), \tag{1}$$

where $S_u \doteq \{\langle x, \lambda_x \rangle \in S \mid \alpha(x) = u\}$. The entropy of a set S is given by

$$\mathrm{ent}(S) \doteq \sum_{\lambda \in \mathcal{L}} -q_\lambda \cdot \log_2(q_\lambda), \tag{2}$$

where $q_\lambda \doteq \mathrm{card}(\{\langle x, \lambda_x \rangle \in S \mid \lambda_x = \lambda\}) \cdot \mathrm{card}(S)^{-1}$. Besides, a number of other selection measures have been devised. See [11] for an empirical comparison of such measures.

Since decision tree induction is by now a well-known method, we shall restrict ourselves to a concise exposition of the basic algorithm underlying ID3 and C4.5. This algorithm derives a decision tree in a top-down fashion by means of the following heuristic (greedy) strategy:

- The complete set of training samples, S, is assigned to the root of the tree.
- A node η becomes a leaf (answer node) of the tree if all associated samples S_η belong to the same class λ. In this case, η is assigned the label λ.[2]
- Otherwise, node η becomes a decision node: It is split by partitioning the associated set S_η of examples. This is done by selecting an attribute (among those that have not been used so far) as described above and by classifying the samples $x \in S_\eta$ according to the values $\alpha(x)$. Each element of the resulting partition defines one successor node.

Once the decision tree has been constructed, each path can be considered as a rule. The antecedent of a rule is a conjunction of conditions of the form $\alpha_\imath(x) = u_\imath$, where $\alpha_\imath$ is an attribute and $u_\imath \in \mathrm{dom}(\alpha_\imath)$ a specific value thereof. The conclusion part determines a value for the class variable. New examples are then classified on the basis of these rules, i.e. by looking at the class label of the leaf node whose attribute values match the description of the example.

3 A Possibilistic Version of Ockham's Razor

3.1 Possibility Theory

Here we briefly review some aspects of possibility theory without going into technical detail. Possibility theory [7] is an alternative calculus for modeling and processing uncertainty or, more generally, partial belief. Possibility theory makes a distinction between the concepts of *certainty* (necessity) and *plausibility* (possibility) of an event. As opposed to probability theory, it does not claim that the confidence in an event is determined by the confidence in the complement of that event. Consequently, possibility theory is non-additive. In fact, the basic axiom of possibility theory involves the maximum-operator rather than the arithmetic sum: $\Pi(A \cup B) = \max\big\{\Pi(A), \Pi(B)\big\}$. In plain words, the possibility of the union (disjunction) of two events A and B is the maximum of the respective possibility of the individual events. A possibility distribution Π on 2^X (satisfying $\Pi(X) = 1$ and $\Pi(\emptyset) = 0$) is related to a possibility measure $\pi : X \to V$ via $\Pi(A) \doteq \sup_{x \in A} \pi(x)$. V is a totally ordered scale which is usually taken as the unit interval $[0, 1]$. However, V can also be a purely qualitative scale, in which case $\pi(x) < \pi(y)$ simply means that y is more plausible than x. A so-called *necessity measure* N, defined by $N(A) \doteq 1 - \sup_{x \in X \setminus A} \pi(x)$ for all $A \subseteq X$, is associated with a possibility measure Π. A necessity measure satisfies $N(A \cap B) = \min\big\{\Pi(A), \Pi(B)\big\}$.

[2] In the case of noisy data, it may happen that all attributes have already been used along the path from the root of the tree to η, though not all samples have the same label.

Where does a possibility distribution come from? Originally, the idea of Zadeh [14] was to induce a possibility distribution from vague linguistic information, as represented by a fuzzy set. For example, the uncertainty related to the vague statement that "x is a small positive integer" translates into a distribution which lets $x = 1$ appear fully plausible ($\pi(1) = 1$), whereas, say, 5 is regarded as more or less plausible ($\pi(5) = 1/2$) and 10 as impossible ($\pi(10) = 0$).[3]

More generally, a possibility distribution can be induced by a *flexible constraint*: Consider a set A of alternatives and suppose information about an element $a_0 \in A$ of interest to be given, expressed in the form of a constraint. Usually, a constraint completely excludes some alternatives $a \in A$ and can hence be identified with a subset $C \subseteq A$ of still admissible candidates. A *flexible* constraint may exclude alternatives to a certain extent. A possibility degree $\pi(a)$ is then understood as the plausibility that remains of alternative a given the constraint.

Note that two constraints are naturally combined by intersection. The possibilistic counterpart to this kind of conjunctive operation is the (pointwise) minimum, i.e. the combination of two possibility distributions π_1 and π_2 into a new distribution $\pi : x \mapsto \min\{\pi_1(x), \pi_2(x)\}$.

In the following section, we shall look at Ockham's razor as a flexible constraint. More generally, our view of a heuristic inductive reasoning principle is that of a constraint which may exclude a model (from the class of candidate models) to a certain degree.

3.2 Ockham's Possibilistic Razor

According to Ockham's razor, a simple model is to be preferred to a more complex one. In the context of decision trees, simplicity is usually equated with size and, hence, one tries to find the smallest tree among those consistent with the data. Note that the heuristic divide and conquer algorithm outlined in Section 2 only finds an approximation to this tree.

Of course, what we actually desire is the *true* model, and the assumption underlying Ockham's razor is that a simple model is more likely to be true than a complex one if both explain the data equally well. Even though this assumption is not very well settled from a theoretical point of view it is intuitively appealing and has proved its worth in practice [5].

Now, consider two decision trees τ^* and τ, where τ is only slightly more complex than τ^*. In such a case, one would generally not completely reject τ. Indeed, when taking the "more likely to" in the above formulation of Ockham's razor seriously, then τ should be assigned a certain degree of possibility as well. This, in turn, should be taken into account when making inferences about new objects. More generally, this *possibilistic* interpretation of Ockham's razor suggests to define a possibility distribution $\pi_{\mathcal{M}}$ over the class of models $\mathcal{M}$, where the possibility $\pi_{\mathcal{M}}(\tau)$ depends on the simplicity of τ in comparison to the simplicity of

[3] The specific definition of π clearly depends on the context.

the simplest (and hence most plausible[4]) model τ^*:

$$\pi_{\mathcal{M}}(\tau) = \pi_{\mathcal{M}}(\tau \,|\, S) \doteq \begin{cases} 0 & \text{if } \tau \text{ is not consistent} \\ f(|\tau|, |\tau^*|) & \text{otherwise} \end{cases}, \tag{3}$$

where $|\tau|$ denotes the complexity of τ (a model τ is consistent if $\tau(x) = \lambda_x$ for all instances $\langle x, \lambda_x \rangle \in S$). A possibilistic prediction, that is a possibility distribution over the class of labels $\mathcal{L}$, can then be obtained by applying the well-known *extension principle*:

$$\pi_{\mathcal{L}}(\lambda) = \pi_{\mathcal{L}}(\lambda \,|\, x_0) \doteq \sup\{\pi_{\mathcal{M}}(\tau) \,|\, \tau(x_0) = \lambda\}. \tag{4}$$

Needless to say, the computation of the possibility measure (3) is generally not tractable, as it requires the consideration of all (consistent) models. Apart from that, one will often not be interested in the possibility degrees of all models, but only in those models with a high degree of possibility. In the following section, we shall propose a heuristic approach which is a generalization of recursive partitioning: The problem of inducing a decision tree is decomposed into sub-problems in a hierarchical way, and the possibility of a tree τ is derived from the possibilities of its sub-trees.

4 Generalized Decision Tree Learning

Recall that the selection of an attribute in decision tree learning is made on the basis of a measure such as (1). Now, suppose that $G(S_\eta, \alpha^*)$ is quite large for the apparently optimal attribute α^*, whereas $G(S_\eta, \alpha)$ is rather small for all remaining attributes. Taking the adequacy of the decision tree approach for granted, one can then be quite sure that α^* is indeed the "correct" selection (problem decomposition) at this place. However, if $G(S_\eta, \alpha)$ is close to $G(S_\eta, \alpha^*)$ for some alternative attribute α, it is reasonable to say that α appears possible to a certain extent as well. More specifically, one might define a degree of possibility $\pi_{\mathcal{A}}(\alpha \,|\, S_\eta)$ for each attribute α on the basis of the set of measures $\{G(S_\eta, \alpha) \,|\, \alpha \in \mathcal{A}\}$, for example

$$\pi_{\mathcal{A}}(\alpha) = \pi_{\mathcal{A}}(\alpha \,|\, S_\eta) \doteq \max\left\{0, 1 - c\left(G(S_\eta, \alpha^*) - G(S_\eta, \alpha)\right)\right\}, \tag{5}$$

where $c > 0$. In order to guarantee a meaningful interpretation of the difference $G(S_\eta, \alpha^*) - G(S_\eta, \alpha)$, the measure $G(\cdot)$ is assumed to be normalized such that $0 \leq G(\cdot) \leq 1$, with 1 being the best evaluation.

This idea suggests the following generalization of the algorithm for decision tree induction: At a node η, a recursive partitioning is not only made for the best attribute a^* but rather for all attributes in the set

$$\mathcal{A}^*_\eta \doteq \{\alpha \in \mathcal{A}_\eta \,|\, \pi_{\mathcal{A}}(\alpha) > \Delta\} \tag{6}$$

[4] Letting $\pi_{\mathcal{M}}(\tau^*) = 1$ for a least one $\tau^* \in \mathcal{M}$ means that at least one model is fully plausible. This can be seen as a kind of closed world assumption. More generally, one might allow that $\pi_{\mathcal{M}}(\tau) < 1$ for all $\tau \in \mathcal{M}$, suggesting that none of the candidate models is fully plausible.

of candidates whose possibility exceeds a lower threshold Δ. More precisely, a *possibilistic branching* is realized as follows: For each attribute $\alpha \in \mathcal{A}^*_\eta$ and each value $u \in \mathrm{dom}(\alpha)$, one outgoing edge is added to η. This edge is marked with the test $\alpha = u$ and the possibility degree $\pi_{\mathcal{A}}(\alpha)$. Thus, one obtains a possibilistic tree or, say, a *meta-tree* T in which an instance can branch at a node in different directions. T actually consists of several ordinary trees τ. In fact, an ordinary tree is obtained by retaining at each (meta-)node η only those edges associated with a single attribute and by deleting all other edges. The possibility of a tree, $\pi_{\mathcal{M}}(\tau)$, is determined by the smallest possibility of its edges.

4.1 Classification with Possibilistic Trees

Now, suppose that a new query x_0 is to be classified. Given the possibility distribution $\pi_{\mathcal{M}}(\cdot)$ as defined above, a possibilistic prediction of the label λ_{x_0} can be derived from (4). However, a more efficient approach is to propagate possibility degrees in the meta-tree T directly. To this end, define possibility distributions $\pi^\eta_{\mathcal{L}}$ for nodes η in a recursive way as follows: If η is a leaf node, then $\pi^\eta_{\mathcal{L}}$ is defined by

$$\pi^\eta_{\mathcal{L}} : \lambda \mapsto \begin{cases} 1 & \text{if } \eta \text{ is labeled with } \lambda \\ 0 & \text{otherwise} \end{cases} .$$

Otherwise, let $\eta_1, \ldots, \eta_r$ be the successor nodes of η, and suppose the edge leading from η to $\eta_\imath$ be marked with the possibility degree $p_\imath$. The distribution associated with η is then given by

$$\pi_{\mathcal{L}} : \lambda \mapsto \max_{1 \leq \imath \leq r} \min\{\pi^{\eta_\imath}_{\mathcal{L}}(\lambda), p_\imath\}. \tag{7}$$

The possibility distribution $\pi_{\mathcal{L}} = \pi_{\mathcal{L}}(\cdot \,|\, x_0)$ is defined to be the possibility distribution $\pi^{\eta_0}_{\mathcal{L}}$ associated with the root η_0 of the meta-tree.

Proposition 1. *The propagation of possibility degrees in the meta-tree yields the same possibilistic prediction $\pi_{\mathcal{L}}(\cdot \,|\, x_0)$ as the extension principle (4).*

Proof. Let $\pi_{\mathcal{L}}$ be the possibility distribution derived from the propagation of possibility degrees in the meta-tree T. Moreover, consider a label $\lambda \in \mathcal{L}$ and let $p = \pi_{\mathcal{L}}(\lambda)$. If $p = 0$ then none of the leaf nodes in T is labeled with λ, and the proposition is obviously correct. Now, let $p > 0$. The definition (7) of distributions associated with nodes entails the existence of a path $\rho^* = (\eta_1, \ldots, \eta_k)$ in T such that the following holds: (1) η_1 is the root of τ and η_k is a leaf node with label λ. (2) The possibility $\pi(\rho^*)$ of the path ρ^*, that is the minimum of the possibility degrees assigned to the edges $(\eta_\imath, \eta_{\imath+1})$, $1 \leq \imath < k$, is given by p. Moreover, $\pi(\rho) \leq p$ for all other paths ρ in the meta-tree whose leaf nodes are labeled with λ.

Now, it is easily verified that the path ρ^* can be completed to an ordinary decision tree τ such that $\pi_{\mathcal{M}}(\tau) = d$. In fact, at each node η in the meta-tree T there is an attribute α such that all edges associated with that attribute are

labeled with the possibility degree 1. Thus, the path ρ^* can be extended to a tree τ such that each edges of τ which is not an edge of ρ^* is labeled with a possibility degree of 1. Therefore, $\pi_{\mathcal{M}}(\tau) = p$, which means that the possibility of λ according to (4) is at least p. Clearly, (4) cannot be larger than p, since this would imply the existence of a tree τ which assigns x_0 the label λ and whose edges all have possibility degrees larger than d. This tree therefore contains a path ρ whose leaf node is labeled with λ and such that $\pi(\rho) > p$, a contradiction to the definition of ρ^*. Therefore, the possibility of λ according to (4) is also given by d. □

Using the classification scheme outlined above, a single estimated class label λ_0 as predicted by an ordinary decision tree is replaced by a prediction in the form of a possibility distribution $\pi_{\mathcal{L}}$ over the set of labels. This distribution is normalized in the sense that $\max_{\lambda \in \mathcal{L}} \pi_{\mathcal{L}}(\lambda) = 1$. Note that the label λ_0^* with $\pi_{\mathcal{A}}(\lambda_0^*) = 1$ is unique unless there is an exact equivalence $G(S_\eta, \alpha_\imath) = G(S_\eta, \alpha_\jmath)$ for a node η and two attributes $\alpha_\imath \neq \alpha_\jmath$. If λ_0^* is unique, it is just the label predicted by the classical approach to decision tree induction.

The distribution $\pi_{\mathcal{A}}$ reflects the uncertainty related to the classification: λ_0^* is the most plausible classification and will generally be chosen if a definite decision must be made. However, there might be further possible candidates as well, and the related possibility degrees indicate the reliability of λ_0^*. Formally, reliability is reflected by the *necessity* degree of λ_0, given by $1 - \max_{\lambda \neq \lambda_0^*} \pi_{\mathcal{L}}(\lambda)$: If there is at least one other label with a rather high degree of possibility, the situation is ambiguous. A classification (on the basis of a decision tree) might then be rejected. More generally, one might take action on the basis of a set-valued prediction including the maximally plausible labels, or take this set as a point of departure for the acquisition of further information.

The approach proposed here is related to other extensions of decision tree learning. Especially, the idea of *option decision trees* [3,9], which also provide a compact representation of a class of candidate decision trees, is worth mentioning in this connection. There are, however, some important differences between the two methods. For example, the outcomes at an option node are combined to a unique choice, e.g. by means of a majority vote. As opposed to this, our approach considers different choices with different degrees of possibility.

4.2 Alternative Aggregation Procedures

Consider a meta-tree T and let $\mathcal{P} = \mathcal{P}_{x_0}$ denote the class of paths ρ in T that are matched by the new query x_0 (where x_0 matches a path if it satisfies all tests $\alpha_\imath(x_0) = u_\imath$ along that path). In agreement with the common max-min calculus of possibility theory we have defined the possibility of a path $\rho = (\eta_1, \ldots, \eta_k)$ as

$$\pi_{\mathcal{P}}(\rho) \doteq \min_{1 \leq \imath < |\rho|} \text{poss}((\eta_\imath, \eta_{\imath+1})), \tag{8}$$

where $\text{poss}((\eta_\imath, \eta_{\imath+1}))$ denotes the possibility degree assigned to the edge $(\eta_\imath, \eta_{\imath+1})$. Moreover, the possibility of a label λ was determined as

$$\pi_{\mathcal{L}}(\lambda) \doteq \max_{\rho \in \mathcal{P} \,:\, l(\rho)=\lambda} \pi_{\mathcal{P}}(\rho), \tag{9}$$

where $l(\rho)$ is the label of ρ's leaf node ($\max \emptyset = 0$ by definition). The minimum in (8) and the maximum in (9) are special types of aggregation operators. In fact, the minimum actually serves as a kind of conjunctive aggregation function, whereas the maximum is a special type of disjunctive operator.

These aggregation functions can be replaced by more general operators, namely by a generalized (logical) conjunction, called a t-norm, and a generalized disjunction called a t-conorm. A t-norm is a binary operator $\otimes : [0,1]^2 \to [0,1]$ which is commutative, associative, monotone increasing in both arguments and which satisfies the boundary conditions $x \otimes 0 = 0$ and $x \otimes 1 = x$. An associated t-conorm is defined by the mapping $(\alpha, \beta) \mapsto 1 - (1-\alpha) \otimes (1-\beta)$. As can be seen, $\otimes = \min$ is a special t-norm with associated t-conorm $\oplus = \max$. Other important operators include the product $\otimes_P : (\alpha, \beta) \mapsto \alpha\beta$ with related t-conorm $\oplus_P : (\alpha, \beta) \mapsto \alpha + \beta - \alpha\beta$ and the Lukasievicz t-norm $\otimes_L : (\alpha, \beta) \mapsto \max\{0, \alpha + \beta - 1\}$ with related t-conorm $\oplus_L : (\alpha, \beta) \mapsto \min\{1, \alpha + \beta\}$.

Replacing min and max by a t-norm $\otimes$ and a t-conorm $\oplus$ yields

$$\pi_{\mathcal{P}}(\rho) \doteq \bigotimes_{1 \leq \imath < |\rho|} \text{poss}((\eta_\imath, \eta_{\imath+1})), \tag{10}$$

$$\pi_{\mathcal{L}}(\lambda) \doteq \bigoplus_{\rho \in \mathcal{P} \,:\, l(\rho)=\lambda} \pi_{\mathcal{P}}(\rho). \tag{11}$$

As opposed to max and min, which are in agreement with the interpretation of possibility distributions as generalized constraints, most other operators are *compensatory*. For example, the possibility of a path is completely determined by its weakest edge according to (8), whereas several strong edges might compensate for this edge when using (10). Likewise, a label supported by several moderately possible paths might be preferred to a label supported by one very plausible path when using an operator such as the probabilistic sum $\oplus_P : (\alpha, \beta) \mapsto \alpha + \beta - \alpha\beta$.

Note that the label λ_0^* estimated by an ordinary decision tree τ will always have a possibility degree of $\pi_{\mathcal{A}}(\lambda_0^*) = 1$ in the possibilistic extension. In fact, the path ρ in τ which is matched by x_0 has a possibility degree of 1 in the meta-tree T. Thus, $\pi_{\mathcal{A}}(\lambda_0^*) = 1$ follows immediately from $\alpha \oplus 1 = 1$ which holds true for every t-conorm $\oplus$ and all $0 \leq \alpha \leq 1$. Now, however, it may happen that a label λ is also regarded as fully possible, even though there is no completely plausible path (classification sequence) that yields λ as a label. For example, suppose λ to be supported by at least $\jmath$ paths ρ with possibility $\pi_{\mathcal{P}}(\rho) \geq 1/\jmath$. When using the Lukasievicz t-conorm as an aggregation, one then obtains $\pi_{\mathcal{A}}(\lambda) = 1$.

5 Lazy Decision Tree Learning

Needless to say, a generalized (possibilistic) decision tree T as outlined in Section 4 can become quite awkward. In fact, passing from an ordinary tree to a

possibilistic tree might easily result in a doubling or trebling of the (average) branching factor.

In this connection, the idea of a *lazy* decision tree learner as outlined in [8] is quite interesting: In classical decision tree learning, test attributes are chosen so as to minimize the *average* impurity of the children of a node, thereby supporting the overall objective of maximizing average performance. However, a decision tree thus induced might not be optimally adapted to a specific query x_0. For example, the entropy of the child relevant for x_0 might well increase, even though the average entropy decreases. Lazy decision tree induction applies the idea of lazy learning [1] to decision trees. Roughly speaking, only a single path of an imaginary decision tree is generated, namely the path which is matched by the query x_0. This allows for selecting the test attributes in a manner which is most favorable for the specific instance x_0.[5]

More precisely, the method proposed in [8] – called LazyDT by the authors – works as follows: As usual, the complete set of training samples, S, is assigned to the root of the tree. A node η becomes a leaf (answer node) if all associated samples S_η belong to the same class or if all attributes have already been used along the path from the root to η. Otherwise, the sample S_η associated with η is split according to the values of an attribute. As an evaluation measure for attributes α, a modified version G^* of the information gain (1) is proposed: Firstly, G^* is computed for the sub-sample S_u with $u = \alpha(x)$ alone, not as a weighted average over all sub-samples. Secondly, the instances at a node η are weighted such that each class has equal weight, which means that the parent node has maximal entropy (see [8] for a justification of this approach). Once having identified an optimal attribute α^*, the procedure is called recursively for the sub-sample $S_{\alpha^*(x_0)}$.

Apart from conceptual advantages in comparison to classical decision tree learning, this approach is interesting in our context since it avoids the generation of a complete (meta-)tree: Even though the individual path generated by the lazy learner becomes a "possibilistic path", that is an ordinary tree, within our approach, it can be handled much more efficiently than a meta-tree.

The possibilistic version of LazyDT – call it PLazyDT – performs in the same way as the original approach, with the following exceptions: At a node η, a degree of possibility $\pi_{\mathcal{A}}(\alpha)$ is derived for all (still available) attributes α. This is done as in Section 4, using a normalized version of the G^* measure. Then, one successor node η_α is defined for each attribute $\alpha \in \mathcal{A}^*_\eta$. The sub-sample assigned to η_α is the set of samples $\langle x, \lambda_x \rangle \in S_\eta$ such that $\alpha(x) = \alpha(x_0)$. While generating a path ρ, the possibility degrees $\pi_{\mathcal{A}}(\alpha)$ (assigned to edges of that path) are accumulated using the minimum operator or, more generally, a t-norm as proposed in Section 4.2. When reaching the leaf node of ρ, one thus obtains a predicted label $l(\rho)$ along with a possibility degree $\pi_{\mathcal{P}}(\rho)$. Finally, the possibility $\pi_{\mathcal{L}}(\lambda)$ of a label λ is obtained by combining the possibility degrees of all paths ρ with $l(\rho) = \lambda$, using the maximum or an alternative t-conorm.

[5] Note that a lazy learner needs to store all observations.

6 Experimental Results

As already explained above, the label estimated by an ordinary decision tree is also fully supported by the possibilistic generalization. Thus, the two approaches will principally yield the same final decisions. A difference can only occur if the distribution $\pi_{\mathcal{A}}$ assigns full support to *several* labels. We shall turn to this aspect in Section 6.3 below. Still, the main motivation underlying the possibilistic approach is the idea of indicating the *uncertainty* related to a decision. This point will be investigated in Section 6.2.

In this section, we restrict ourselves to the lazy versions of decision tree induction, as we obtained quite similar results for the classical approaches (apart from the runtime of the algorithms).

6.1 Experimental Setup: Generation of Synthetic Data

An individual experiment is parameterized by the number of attributes, k, the number of labels, m, the size of the training sample, n, and a complexity parameter γ:

- An underlying "true" decision tree τ is generated at random. This is done in a recursive manner by starting with the root of the tree and flipping a (biased) coin to decide whether the current node is an inner node or a leaf.[6] The probability of a node to become an inner node is specified by a fixed parameter $0 < \gamma < 1$ (the larger γ, the more complex the tree will be on average). Here, we restrict ourselves to binary trees, i.e. we only consider binary attributes. Once a leaf node has been generated, it is assigned a class label at random.[7] Likewise, inner nodes are assigned attributes.
- A random sample is generated based on a uniform distribution over the instance space. The sample is labeled using the decision tree τ.
- Decision trees τ_1 and τ_2 are induced, respectively, by LazyDT and PLazyDT based on the random sample.
- A new query x_0 is generated at random and classified by the two trees, which yields an estimation $\lambda_1^* = \tau_1(x_0)$ and a possibilistic prediction $\pi_{\mathcal{A}}$ with related decision $\lambda_2^* = \arg\max_{\lambda \in \mathcal{L}} \pi_{\mathcal{A}}(\lambda)$. The correct label is $\lambda_{x_0} = \tau(x_0)$.

6.2 Representation of Uncertainty

To capture the aspect of uncertainty representation, let p_1 denote the expected degree of possibility assigned by $\pi_{\mathcal{L}}$ to the *correct* label λ_{x_0} given that this label is *not* predicted ($\lambda_2^* \neq \lambda_{x_0}$). Moreover, let p_2 denote the expected possibility of the most possible incorrect label $\lambda \neq \lambda_{x_0}$ given that the decision is correct, that is 1 minus the degree of necessity of λ_{x_0}. Ideally, p_1 is large and p_2 is small: Wrong decisions are accompanied by a large degree of uncertainty, reflected by

[6] The root is never a leaf.

[7] We pay attention that not all successors of a node do have the same label.

considerable support of the actually correct label (and hence a low degree of necessity for λ_2^*), whereas correct decisions appear reliable at the same time.

We have derived approximations to these expected values by taking averages over 10,000 experiments. The table below shows results (r denotes the classification rate) for different setups with $k = 6$, $\gamma = 0.8$. For PLazyDT we used max and min as aggregation operators, the function (5) with $c = 1/3$ for assigning basic possibility degrees, and the threshold $\Delta = 0$ in (6).

	$m = 2$			$m = 3$			$m = 4$		
n	r	p_1	p_2	r	p_1	p_2	r	p_1	p_2
10	0.720	0.700	0.423	0.632	0.653	0.381	0.581	0.533	0.249
20	0.784	0.810	0.363	0.742	0.754	0.336	0.649	0.717	0.244
30	0.838	0.871	0.331	0.762	0.814	0.261	0.734	0.726	0.192
40	0.855	0.886	0.265	0.799	0.839	0.214	0.786	0.767	0.162

As can be seen, the reliability of a prediction is reflected extremely well by the possibilistic estimation. As it was to be expected, both the classification rate and the quality of the possibility distribution (as indicated by p_1, p_2) increase with sample size (as already explained above, the larger p_1 and the smaller p_2, the better the quality of the distribution). For other setups (values k, γ) the results were qualitatively very similar. We do not present them here for reasons of space.

6.3 Classification Performance

One may obtain $\pi_{\mathcal{A}}(\lambda) = 1$ for several labels $\lambda \in \mathcal{L}$ when making use of more general t-norms and t-conorms. In such a case, there are different options to make a final decision. Here, we simply choose one among these labels at random. The following results were again derived for $k = 6, \gamma = 0.8$, using the t-norm $(\alpha, \beta) \mapsto \alpha\beta$ and the related t-conorm $(\alpha, \beta) \mapsto \alpha + \beta - \alpha\beta$ (r_1 and r_2 denote the classification rate for LazyDT and PLazyDT, respectively).

	$m = 2$		$m = 3$		$m = 4$	
n	r_2	r_1	r_2	r_1	r_2	r_1
10	0.731	0.707	0.646	0.652	0.587	0.556
20	0.797	0.792	0.711	0.708	0.675	0.656
30	0.851	0.812	0.760	0.757	0.734	0.730
40	0.864	0.844	0.808	0.782	0.789	0.776

As can be seen, PLazyDT is slightly superior, though – as it was to be expected – the difference in classification performance is not very significant. We obtained quite similar results for several real-world data sets from the UCI repository which are, again for reasons of space, not presented here. These results confirm that aggregating over possible models might indeed be better than completely relying on the supposedly optimal one.

7 Concluding Remarks

Inductive reasoning based on OCKHAM's razor or, more generally, on heuristic principles is always afflicted with uncertainty. The major concern of the method proposed in this paper is to capture this type of uncertainty, which appears to be non-probabilistic by nature. Therefore, our formalization employs the alternative framework of possibility theory (flexible constraints). Let us mention that a related possibilistic formalization has already been developed for the heuristic principle underlying instance-based learning [6].

Of course, one might deplore the lacking of a sound theoretical basis for the possibilistic approach. It should be noted, however, that the same remark already applies to the underlying heuristic principle itself. In fact, what we introduced here is an *alternative formalization* of OCKHAM's razor which – according to our opinion – extends the original version in a reasonable way. As the experimental results confirm, the possibilistic approach represents the reliability of a prediction in a thorough way and may even (slightly) improve classification performance.

Apart from the uncertainty connected to inductive inference one usually has to cope with other types of uncertainty as well, such as e.g. noisy data. Extending the method proposed here by combining these different types of uncertainty is one of the challenges for future work.

References

1. D. W. Aha, editor. *Lazy Learning*. Kluwer Academic Publ., 1997. 181
2. L. Breiman, J. Friedman, R. Olshen, and C. Stone. *Classification and Regression Trees*. Wadsworth International Group, Belmont, CA, 1984. 174
3. W. Buntime. Learning classification trees. *Statistics and Computing*, 2(2), 1992. 179
4. L. J. Cohen. *An Introduction to the Philosophy of Induction and Probability*. Claredon Press, Oxford, 1989. 173
5. P. Domingos. The role of Occam's razor in knowledge discovery. *Data Mining and Knowledge Discovery*, 3:409–425, 1999. 176
6. D. Dubois, E. Hüllermeier, and H. Prade. Fuzzy set-based methods in instance-based reasoning. IEEE *Transactions on Fuzzy Systems*. To appear. 184
7. D. Dubois and H. Prade. *Possibility Theory*. Plenum Press, 1988. 175
8. J. H. Friedman, R. Kohavi, and Y. Yun. Lazy decision trees. In *Proceedings* AAAI–96. Morgan Kaufmann, 1096. 181
9. R. Kohavi and C. Kunz. Option decision trees with majority votes. In *Proceedings* ICML–97. 179
10. J. Mingers. An empirical comparison of pruning methods for decision tree induction. *Machine Learning*, 4:227–243, 1989.
11. J. Mingers. An empirical comparison of selection measures for decision-tree induction. *Machine Learning*, 3:319–342, 1989. 174
12. J. R. Quinlan. Discovering rules by induction from large collections of examples. In D. Michie, editor, *Expert Systems in the Micro Electronic Age*. 1979. 174
13. J. R. Quinlan. *C4.5: Programs for Machine Learning*. Morgan Kaufmann, San Mateo, CA, 1993. 174
14. L. A. Zadeh. Fuzzy sets as a basis for a theory of possibility. *Fuzzy Sets and Systems*, 1:3–28, 1978. 176