Growing Decision Trees in the presence of Indistinguishability: Observational Decision Trees.

Enric Hernández (enriche@lsi.upc.es)* Jordi Recasens (recasens@ea.upc.es)†

Keywords: decision tree, T-indistinguishability operators, observational entropy, uncertainty measures, machine learning.

1 Introduction.

Decision trees, since their formal appearance within the context of inductive learning [9] have become one of the most relevant paradigm of machine learning methods. The main reason for this widespreading success lies in their proved applicability to a broad range of problems, in addition to appealing features as the readibility of the knowledge represented in the tree. Therefore, a lot of work have been carried out from Quinlan's TDID3 algorithm in order to extend the applicability to domains beyond the categorical ones and achieve further improvements. In this line, many approaches dealing with continuous-valued attributes have been proposed ([1, 10, 8]). Also, alternative measures to classical Shannon's entropy measure [2] for attribute selection have been devised, like Gini's test [1], Kolmogorov-Smirnoff distance [13], distance between partitions [11], contrast measures [3] ...

Another important point is providing decision tree induction algorithms with a more flexible metodology in order to cope with other sources of uncertainty beyond the probabilistic type. Indeed, when we face real problems we should overcome the limitation of the probabilistic framework by furnishing existing methods, so that other well-known types of uncertainty such as non-specifity and fuzziness [7] could be managed. [6], [14], [15], [12] are worthwhile methods concerning to this problem.

In this paper we will address the case when uncertainty arises as a consequence of having defined an indistinguishability relation [5] on the domains of the attributes used to describe the set of instances. As far as we know, existing methods make the assumption that different events are perfectly distinguishable from each other when measuring, for instance, node's impurity (for entropy-based methods). In front of the above restrictive assumption we advocate for a more realistic setting in which decision maker's discernment abilities should be taken into account, and therefore, impurity should be measured accordingly to his frame of discernment. With this purpose in mind we introduce the notion of observational entropy which adapts the classical definition of entropy in order to incorporate such indistinguishability concerns. The main idea is that the occurrence of two different events but indistinguishable by the indistinguishability relation defined, will count as the occurrence of the same event when measuring the observational entropy.

*Research supported by DGICYT project number PB98-0924

†Secció de Matemàtiques i Informàtica. ETSAB. Avda. Diagonal 649. 08028 Barcelona. Spain. Universitat Politècnica de Catalunya

2 Observational entropy.

In this section we will present the definition of observational entropy and conditioned observational entropy which will be used in later sections.

Definition 1 *Given a t-norm T, a T-indistinguishability operator E on a set X is a reflexive and symmetric fuzzy relation on X such that $T(E(x,y),E(y,z)) \leq E(x,z)$ (T-transitivity), for all $x,y,z \in X$.*

Throughout the paper E and E' will denote T-indistinguishability operators on a given set X and P a probability distribution on X.

Definition 2 *The observation degree of $x_j \in X$ is defined by:*

$$\pi(x_j) = \sum_{x \in X} p(x)E(x,x_j).$$

Due to the reflexivity of E, this expression can be rewritten as:

$$\pi(x_j) = p(x_j) + \sum_{x \in X | x \neq x_j} p(x)E(x,x_j).$$

This definition has a clear interpretation: the possibility of observing x_j is given by the probability that x_j really happens (expressed by the first term), plus the probability of ocurrence of some element "very close" to x_j, weighted by the similarity degree .

Definition 3 *The observational entropy (HO) of the pair (E,P) is defined by:*

$$HO(E,P) = -\sum_{x \in X} p(x)\log_2 \pi(x).$$

The next step is to define the conditioned observational entropy. Informally, the conditioned observational entropy measures how do affect the observations performed by an observer "using" a T-indistinguishability operator E' in the variability degree of the potential observations (observational entropy) of some other observer using another T-indistinguishability operator E.

Definition 4 *$\forall x \in X$ we define:*

$$P^E_{x_j}(x) = \frac{p(x) \cdot E(x, x_j)}{\pi_E(x_j)} = \frac{p(x) \cdot E(x, x_j)}{\sum_{y \in X} p(y) \cdot E(y, x_j)}.$$

That is, $P^E_{x_j}(x)$ quantifies the contribution of x to the observation degree of x_j in (E, P).

Definition 5 *The conditioned observation degree of $x_i \in X$ having been observed x_j in (E', P) as*

$$\pi^{E|E'}_{x_j}(x_i) = \sum_{x \in X} P^{E'}_{x_j}(x) \cdot E(x, x_i).$$

Definition 6 *The observational entropy of the pair (E, P) conditioned to the observation of $x_j \in X$ in (E', P) as follows:*

$$HO_{x_j}(E \mid E', P) = -\sum_{x_i \in X} P^{E'}_{x_j}(x_i) \cdot \log_2 \pi^{E|E'}_{x_j}(x_i).$$

Definition 7 *The observational entropy of the pair (E, P) conditioned by the pair (E', P) as*

$$HO(E \mid E', P) = \sum_{x_j \in X} p(x_j) \cdot HO_{x_j}(E \mid E', P).$$

In other words, the conditioned observational entropy of the pair (E, P) is the expected value of the observational entropy of (E, P) conditioned to the observation of all $x_j \in X$ in (E', P).

3 Algorithm.

In section 2 we introduced the concept of observational entropy. Let us see how to use it for the task of building a decision tree from a set of examples. The problem could be posed as follows: Let $At = \{A_1, \cdots, A_n, C\}$ be a set of nominal[1] attributes (being the classes of C the classification we want to learn), with domains $D_i = \{v_{i_1}, \cdots v_{i_{m_i}}\}$ and $D_c = \{v_{c_1}, \cdots, v_{c_{m_c}}\}$. Let $S \subseteq D_1 \times \cdots \times D_n \times D_c$ be the set of instances, and for each attribute A we consider a T-indistinguishability operator E_A and a probability distribution P_A defined on the domain of A. Let us illustrate the above definitions with the example of tables 1 and 2. In order to simplify, we will assume that the probability distribution associated to each attribute of the example will be defined as the uniform distribution on the corresponding domain. Generalizing this assumption is straightforward.

Arrived at this point, let us present an algorithm for building a decision tree based on the observational entropy. The procedure could be summarized in the following points:

i) "Unfolding" data set: from the original data set we create its associated "unfolded" data set by splitting each column (representing an specific attribute A_i), creating a new column for each value (modality) belonging to the domain of the corresponding attribute. Then, for all instances, we compute the compatibility between each modality and the evidence represented in an instance by computing the conditioned observational degree (5) between the given modality and the proper component (evidence) of the instance. The resulting "unfolded" data set is depicted in table 3.

ii) Computing probabilities of observing events in a node N. Values contained in the unfolded data set will be used to compute the compatibility degree (g) between a conjunction of restrictions and the evidence represented by a given instance s:

$$g(A_i = v_{i_j} \wedge \cdots \wedge A_k = v_{k_l} | s) = t(\pi(v_{i_j} | s.A_i), \cdots, \pi(v_{k_l} | s.A_k))$$

(being T a t-norm) So, being T the current tree (the one which has been grown up to now), N a given node belonging to T and R the conjunction of the restrictions found in the path going from the root of T to node N, we define the probability of observing modality v_{i_j} of attribute A_i in node N as:

$$P_N(A_i = v_{i_j}) = \frac{\sum_{s \in S} g((R \wedge A_i = v_{i_j}) | s)}{\sum_{v_i \in D_i} \sum_{s \in S} g((R \wedge A_i = v_i) | s)}$$

iii) Selecting branching attribute: in the previous point we have provided a method for computing the probabilities of observing the modalities for all the attributes in a given node N. These values will allow us to select the best attribute in order to partition data "arriving" at node N (fulfilling the restrictions leading to node N). In this way, given a node N, we compute (for all non previously selected attributes) the observational entropy of class attribute (C) conditioned to a given remaining attribute A_i in the following manner:

$$HO(C|A_i) = \sum_{v_i \in D_i} P_N(A_i = v_i) \cdot HO(C|A_i = v_i)$$

[1] We consider nominal attributes for simplicity purposes, although the developed metodology can also deal with continuous domains.

,being

$$HO(C|A_i = v_i) = -\sum_{v_c \in D_C} P_{N \wedge (A_i = v_i)}(c = v_c)$$
$$\cdot \log_2 \sum_{w_c in D_C} P_{N \wedge (A_i = v_i)}(C = w_c) \cdot E_C(w_c, v_c)$$

(where $P_{N \wedge (A_i = v_i)}$ are the probabilities measured in each one on the childs of N induced by partition data arriving at node N accordingly with the modalities of attribute A_i)

We select, as current branching attribute, the one which minimizes the conditioned observational entropy (which is equivalent to say that maximizes the observational information gain), and mark it as already used attribute.

iv) Putting all together. Finally we will present the general procedure which, making use of the definitions presented in previous points, is able to induce a decision tree from a set of instances.

1. Create the "unfolded" version of the original data set.
2. Place the initial data on the root.
3. Select the best attribute from the set of non used attributes and mark it as used.
4. Create new child nodes according to the partition induced by the selected attribute.
5. For each newly generated child node iterate step 3 if the following conditions hold:

   - There are remaining non used attributes.

   - The set of instances arriving to that node is not the empty set.

   - Observational entropy of current node is not below a predefined threshold value.

For data in table 1 the induced observational decision tree is:

```
root
|__outlook=sunny
|      |__windy=true
|      |     |__swimming
|      |__windy=false
|            |_volley,tennis
|__outlook=overcast
|      |__windy=true
|      |     |__football
|      |__windy=false
|            |__tennis
|__outlook=rainy
       |__football
```

References

[1] L. Breiman et al. *Classification and regression trees.* Wadsworth International Group, 1984.

[2] Shannon . C. and W. Weaber. *The mathematical theory of communication.* University of Illinois Press, 1964.

[3] Van de Merckt. Decision trees in numerical attribute spaces. In *Proceedings of the 13th International Joint Conference on Artificial Intelligence*, pages 1016–1021, 1993.

[4] Hernandez E. and Recasens J. A reformulation of entropy in the presence of indistinguishability operators. *to appear in Fuzzy sets and Systems.*

[5] Jacas J. and Recasens J. Fuzzy t-transitive realtions: eigenvectors and generators. *Fuzzy sets and systems*, (72):147–154, 1995.

[6] C. Janikow. Fuzzy decision trees: issues and methods. *IEEE Transactions on Systems, Man and Cybernetics*, 28(1):1–14, 1998.

[7] G Klir and M. Wierman. *Uncertainty based information. Elements of generalized information theory.* Physica-Verlag, 1999.

[8] P.E Maher and D. Saint-Clair. Uncertain reasoning in an id3 machine learning framework. In *2nd IEEE Conf. on Fuzzy Systems*, pages 7–12, 1993.

[9] J.R Quinlan. Induction of decision trees. *Machine Learning*, pages 81–106, 1986.

[10] J.R Quinlan. *C4.5: programs for machine learning.* Morgan Kaufmann, 1993.

[11] Mantaras R. A distance-based attribute selection measure for decision tree induction. *Machine learning*, 6(1):81–92, 1991.

[12] M. Umano et al. Fuzzy decision trees by using fuzzy id3 algorithm and its application to diagnosis systems. In *Proceedings 3rd IEEE International Conference on Fuzzy Systems*, pages 2113–2118, 1994.

[13] P.E. Utgoff and J.A. Clouse. A kolmogorov-smirnoff metric for decision tree induction. Technical Report 96-3, University of Massachusets, 1996.

[14] R. Weber. Fuzzy-id3: a class of methods for automatic knowledge acquisition. In *Proceedings 2nd International Conference on Fuzzy Logic and Neural Networks*, pages 265–268, 1992.

[15] Y. Yuan and Shaw M. Induction of fuzzy decision trees. *Fuzzy Sets and Systems*, (69):125–139, 1995.

outlook	*temperature*	*windy*	*play*
sunny	hot	false	volley
sunny	hot	true	swimming
overcast	hot	false	tennis
rainy	mild	false	football
rainy	cool	true	football
overcast	cool	true	football
sunny	mild	false	tennis
sunny	mild	true	swimming
overcast	hot	false	tennis
rainy	mild	true	football

$D_{Outlook} = \{sunny, overcast, rainy\}$
$D_{Temperature} = \{hot, mild, cool\}$
$D_{Windy} = \{true, false\}$
$D_{Play} = \{swimming, tennis, football, volley\}$

Table 1: Original data set.

$E_{Outlook} =$

	sunny	*overcast*	*rainy*
sunny	1	0	0
overcast	0	1	0.5
rainy	0	0.5	1

$E_{Temp} =$

	hot	*mild*	*cool*
hot	1	0.5	0.5
mild	0.5	1	0.5
cool	0.5	0.5	1

$E_{Play} =$

	swimming	*football*	*tennis*	*volley*
swimming	1	0	0	0
football	0	1	0.25	0.25
tennis	0	0.25	1	1
volley	0	0.25	1	1

$E_{Windy} =$

	true	*false*
true	1	0
false	0	1

Table 2: T-Indistinguishability operators (matricial representation).

sunny	*overcast*	*rainy*	*hot*	*mild*	*cool*	*true*	*false*	*swimming*	*tennis*	*football*	*volley*
1	0	0	0.7	0.6	0.6	0	1	0	0.9	0.9	0.1
1	0	0	0.7	0.6	0.6	1	0	1	0	0	0
0	0.8	0.6	0.7	0.6	0.6	0	1	0	0.9	0.9	0.1
0	0.6	0.8	0.6	0.7	0.6	0	1	0	0.5	0.5	0.7
0	0.6	0.8	0.6	0.6	0.7	1	0	0	0.5	0.5	0.7
0	0.8	0.6	0.6	0.6	0.7	1	0	0	0.5	0.5	0.7
1	0	0	0.6	0.7	0.6	0	1	0	0.9	0.9	0.1
1	0	0	0.6	0.7	0.6	1	0	1	0	0	0
0	0.8	0.6	0.7	0.6	0.6	0	1	0	0.9	0.9	0.1
0	0.6	0.8	0.6	0.7	0.6	1	0	0	0.5	0.5	0.7

Table 3: Unfolded data set