Increasing acceptability of decision trees with domain attributes partial orders

Joan Albert López-Vallverdú, David Riaño
Research Group on Artificial Intelligence
Universitat Rovira i Virgili (Tarragona)
{joanalbert.lopez,david.riano}@urv.cat

Antoni Collado
Grup Sagessa
acollado@grupsagessa.com

Abstract

There are several domains, such as health-care, in which the decision process usually has a background knowledge that must be considered. We need to maximize the accuracy of the models, but we also need them to be meaningful. Otherwise it will lead to the problem that the expert finds the obtained models incomprehensible. We propose a way for representing the knowledge of the experts in order to modify the C4.5 algorithm to produce decision trees which are more comprehensible to medical doctors without losing accuracy.

1. Introduction

Decision making is a common activity in medicine. So, for example, diagnosis, drug and therapy prescription, or prognosis are about deciding on the patient disease, treatment, or evolution. In medicine, good decisions are not only those which obtain good results (i.e. accurate decisions) but also those which have a medical sense (i.e. meaningful decisions). Artificial Intelligence has a long tradition in the generation of decisional structures ranging from statistical approaches as Bayesian Networks [1] to symbolic approaches as decision trees [5], or decision tables [9].

In the context of producing decisional structures in medicine, success can be measured at the level of *accuracy* (i.e. is the structure taking good decisions?) or at the level of *meaning* (i.e. has the decision process of the structure a medical sense?). Some of the above mentioned decisional structures as Bayesian Networks have not an explicit representation of the decision process, and therefore measuring the medical meaning of a decision is not possible. Some others are exclusively centered in the construction of accurate structures not necessarily avoiding the generation of medically incomprehensible decision models. One may argue that these models obtained from the evidence on the data may hide decisional aspects that medical doctors may accept and adopt after a deeper analysis, but reality shows that what it normally happens is that medical doctors do not trust these decisions [2].

As far as accuracy is concerned, C4.5 [6] is a successful algorithm for inducing decision trees from instances on the domain in which decisions have to be made. These instances are described in terms of a set of attributes and the values these take. The C4.5 algorithm shares the approach of ID3 [7] that consists on repeatedly select the attribute that partitions the set of instances in a more convenient way, till the parts of the set of instances belong to a single class (or decision alternative). ID3 is based on the *information gain criterion* which measures the quantity of information [8] gained by partitioning the set of instances (also called training set) in accordance with the mutually exclusive outcomes of a single attribute. In each step, the attribute maximizing the gain is the one selected. The main disadvantage of this approach is that it has a strong bias in favor of the attributes with many outcomes. C4.5 solves this by using the

Twentieth IEEE International Symposium on Computer-Based Medical Systems (CBMS'07)
0-7695-2905-4/07 $20.00 © 2007 IEEE

so-called *gain ratio criterion*. This measure includes the concept of split information which penalizes the attributes that produce a wider distribution of the data.

Along the years, multiple works have proved C4.5 to be an efficient machine learning algorithm to generate decision trees, according to the accuracy of the results. However, less works have been published on the analysis of the quality of C4.5 in the generation of meaningful results. Here, we propose the use of partial orders to capture the concept of priority of the attributes in the domain where the decision tree is being induced. In health-care, the priority in the selection of attributes cannot always be represented as a total cost function [4] but as a partial order among the attributes. Experts in medicine can deploy these structures to introduce background knowledge about the target domain in order to guide the C4.5 algorithm in the selection of the most convenient attribute at each step of the learning loop. Now, the convenience of an attribute is a trade off between information gain and domain relevance.

In the next section, the concepts of partial order to represent the relevance of the attributes in the domain, and the evaluation measures of accuracy and meaning are formalized. In section 3, these elements are incorporated in the C4.5 algorithm. The resulting algorithm is PS-C4.5 and it is tested on 6 public medical domains. The results of the tests and their analysis are provided in section 4. Section 5 contains conclusions.

2. Representing and using the medical relevance of attributes

The induction of decision trees is conditioned to the attributes that are used to describe the instances in the training set. Given a concrete set of instances, the information gain of each attribute is a ratio that can be calculated but which do not correspond to the medical sense of the decision process. The medical sense of the attributes must be expressed with a structure establishing the medical preferences among the attributes.

2.1 Partially ordered set of attributes

Provided a set of attributes *A,* the relevance of these attributes in the application domain is represented by a partial order $P \subseteq A \times A$ which is formally described as a binary relation over *A* with the following properties: reflexivity *(* $\forall a \in A$ *aPa),* antisymmetry *(* $\forall a, b \in A$ *if aPb and bPa then a = b),* and transitivity *(* $\forall a, b \in A$ *if aPb and bPc then aPc).*

A set equipped with a partial order relation is called *partially ordered set* or *poset*.

A partial order on the set of attributes of a decision process can be used to represent several meanings. For example, a relationship *age P irradiate* may mean that during a diagnosis procedure of a *breast cancer* it is better to start asking the *age* than asking *irradiate* which represents a traumatic medical procedure. Alternatively, it could also mean that *age* is a better choice because it discards more alternative diagnosis or because there is some medical evidence that a diagnostic process must prioritize asking the *age* than the *irradiate* condition.

2.2 Attribute selection

One of the key aspects of algorithms like C4.5 is their criterion to select the attributes to partition the training set and, therefore, the attributes that are going to be part of the final decision tree. In C4.5, this criterion (*C4.5AttributeSelection*) is based on the attribute information gain. The same way, if a poset is defined on the set of attributes, an alternative criterion for attribute selection is to choose attributes in the order the poset indicates. Let us call this the *PSAttributeSelection* criterion.

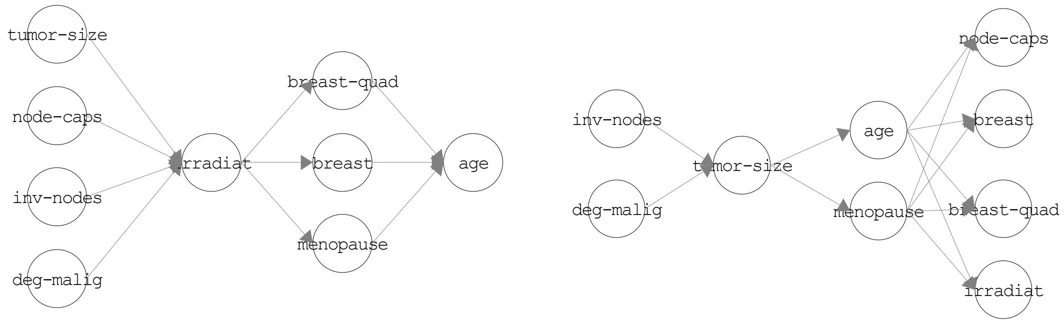

Figure 1. (a) Decision poset equivalent to the C4.5 criterion in the first step; (b) Decision poset provided by the doctor.

In figure 1a we can see the poset equivalent to the first step of the C4.5 process of induction, which is very different from figure 1b which represents the doctor's criterion of selection. Both criteria can be combined to produce alternative policies that may range from pure C4.5 to approaches that only consider the medical meaning embedded in the poset.

2.3 Evaluation measures

Once obtained, it is necessary to evaluate the decision process. Firstly, the model must carry out good decisions. We use the common measure of *accuracy* in equation 1 to quantify how good a decision process is.

$$Accuracy = \frac{Correctly\ classified\ instances}{Number\ of\ instances} \qquad (1)$$

However, this measure is not sufficient because we need to evaluate the comprehensibility of the decision process. We created a measure called *DS (Doctor's Satisfaction)* which scores a decision process in the form of a decision tree according to how well it follows the doctor's criterion of selection. Given a decision tree, the first step consists on transforming it into a poset. Beginning at the root, each level of the tree matches to a level of priority in the poset. The process continues until all the attributes in the tree have been treated. If there are attributes which do not appear in the tree, they are situated in the last level of the poset. Supposing the partial order P_1 provided by the doctor and P_2 obtained from a decision tree on the set of attributes A, we define $A_i = \{(a,b) \in A \times A \mid (a\, P_i\, b)\}$ the set of comparable attribute pairs in P_i. Then the symmetric difference between A_1 and A_2 is $A_1 \Delta A_2 = (A_1 \cup A_2) - (A_1 \cap A_2)$, its cardinality is a measure of how different P_1 and P_2 are, and DS in equation 2 is a measure of the similarity between the poset provided by the doctor and the order used to select the attributes in the creation of the decision tree.

$$DS = 1 - \frac{(card\,(A_1 \Delta A_2))}{(card\ A_1 + card\ A_2)} \qquad (2)$$

3. The PS-C4.5 algorithm

The concepts in the previous section are introduced in the C4.5 basic structure in order to obtain a new algorithm that could improve doctor's satisfaction on the decision trees generated, without a worsening of the global accuracy. The new algorithm is called PS-C4.5. The main difference in comparison with C4.5 is in the selection of the attribute to split the data in each node. In figure 2 we can see a simplified version of this fragment of the algorithm. Firstly, it gets the subset F of attributes with a higher priority (*PSAttributeSelection*). The next step consists on finding the best of the attributes in the subset *F* in accordance to the *C4.5AttributeSelection* criterion. As we have mentioned in section 2.2, there can be several combinations between *PS* and *C4.5AttributeSelection*. We have implemented this by using an information gain threshold called δ in the algorithm. The information gained by the selected attribute must be greater or equal than δ. If δ=0 there will be no constraints and the partial order will be strictly respected. Another possibility is to set δ to the average value of the information gain. In this case, there will be more balance between both criteria. If it finds an attribute whose information gain is greater or equal than δ, the algorithm returns the split model based on this attribute. Otherwise, it removes all the attributes in *F* from the poset and begins again getting the attributes with the highest priority in the poset and looking for the best one.

```
input : P : partial order over the set of attributes A
        I : Instances
output : C4.5 split model for the selected attribute
________________________________________________

$F \leftarrow PSAttributeSelection(P, A);$
$bestAtt \leftarrow \varnothing;$
while $bestAtt = \varnothing \wedge notEmpty(P)$ do
  $bestAtt \leftarrow C4.5AttributeSelection(F);$
  if $infoGain(bestAtt) < \delta$ then
    Remove all the attributes in F from the poset ;
    $F \leftarrow PSAttributeSelection(P, A);$
    $bestAtt \leftarrow \varnothing;$
  end
end
return model(bestAtt);
```

Figure 2. Getting the split model for the chosen attribute according to PS-C4.5 (simplified)

4. Tests and results

To verify the functioning of PS-C4.5 we have applied it in some health-care domains. We have chosen six medical domains [3] and designed a poset for each of them with the information about medical relevance among the attributes. Table 1 contains a brief summary about the main characteristics of each domain.

Table 1. Summary table of problems tested

	Instances	Attributes	Classes	Missing attr.?
HEART DISEASE	920	13	2	Yes
HEPATITIS	155	19	2	Yes
BUPA LIVER DISORDERS	345	6	2	No
PIMA INDIANS DIABETES	768	8	2	No
ECHOCARDIOGRAM	132	6	2	Yes
BREAST CANCER	699	9	2	Yes

For each medical domain we have produced decision trees with C4.5 (*C4.5AttributeSelection* criterion), PS-C4.5 with δ=0 (*PSAttributeSelection* criterion) and δ equal to the average value of the attributes information gain (balance between both criteria).

So, for breast cancer, figure 3 depicts the decision tree we obtained using C4.5 and figure 4 the tree built using PS-C4.5 with δ=0 according to the partial order in figure 1b. We observe in the second one that the attributes are used in a way which is more similar to the order represented by the poset (see figure 1.b) than the C4.5 tree, so it is medically more coherent.

This is also perceived in table 2 with the values of DS, always greater in poset-based trees than in C4.5 trees.

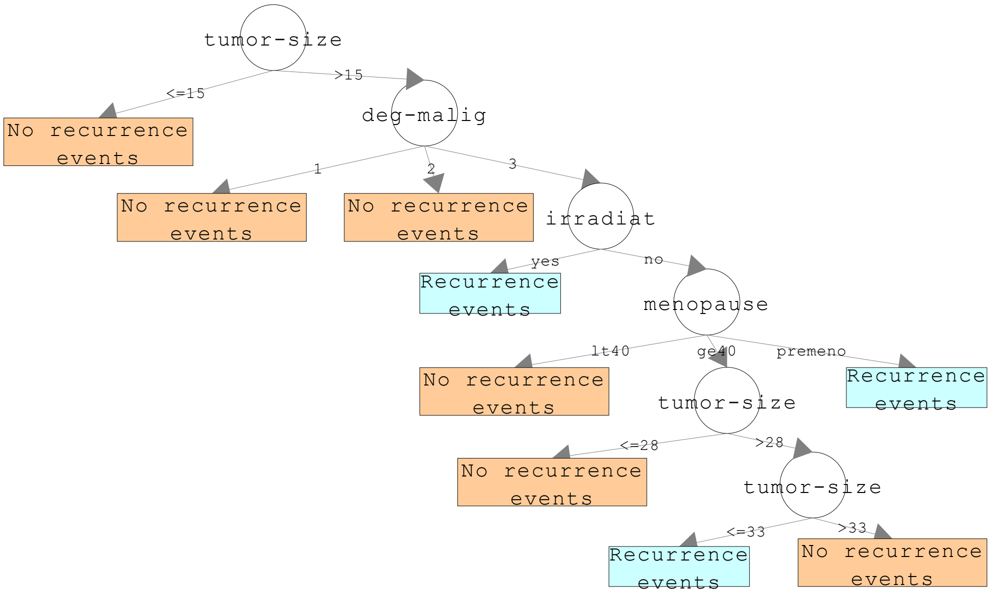

Figure 3. Decision tree induced by C4.5 in the breast cancer problem

With the equations introduced in section 2.3, we evaluated the accuracy and the doctor's satisfaction using a 10-fold cross validation for all the medical domains (see table 2). We point out that PS-C4.5 with δ=0 achieves the best results according to DS measure in each domain. As expected, the mean DS increases as the poset acquires more relevance in the algorithm, reaching 12,2% in average. Using δ=0 causes an average 9,3% DS improvement with respect to δ=avgInfoGain.

Table 2. Results obtained by each algorithm

	C4.5		PS-C4.5 δ=avgInfoGain		PS-C4.5 δ=0	
	Accuracy	DS	Accuracy	DS	Accuracy	DS
HEART DISEASE	0,7657	0,7190	0,7954	0,6541	0,7888	0,8533
HEPATITIS	0,8387	0,4051	0,8065	0,4796	0,8452	0,6164
BUPA LIVER DISORDERS	0,6870	0,4615	0,6493	0,4800	0,6493	0,4800
PIMA INDIANS DIABETES	0,7383	0,3721	0,7240	0,4103	0,7174	0,4103
ECHOCARDIOGRAM	0,7162	0,5000	0,7162	0,5000	0,7703	0,6667
BREAST CANCER	0,7413	0,6667	0,7098	0,7778	0,7238	0,8312
MEAN	0,7479	0,5207	0,7335	0,5503	0,7491	0,6430

It is also interesting to notice that the lost of accuracy of PS-C4.5 with respect to C4.5 is always below 4% but in average it is slightly better for PS-C4.5 δ=0. When it is observed, this lost of accuracy is caused by the pruning algorithm that C4.5 applies and not by the tree induction process directly.

5. Conclusions

We have carried out an approach to the representation of the medical knowledge applied to decision problems. From a medical point of view it has been contrasted that, although the trees obtained with C4.5 are correct and in some cases they are comprehensible enough, the trees created by PS-C4.5 are preferred by physicians because they seem more logical and consistent in accordance to the standards of medical knowledge and practice. The measures used to test the models show that the poset-based trees are better than C4.5 in accordance to DS. And it is important to highlight that the accuracy obtained by PS-C4.5 is always equivalent (less than 4% different) to the accuracy of the C4.5 trees.

We have also noticed that the priority of selecting an attribute in a process of decision can vary over time. The previous selection of an attribute can increase or decrease the priority of another attribute. Thus, we plan the possibility to improve the algorithm in the future by adding dynamic posets which change their structure along the process of induction of the tree.

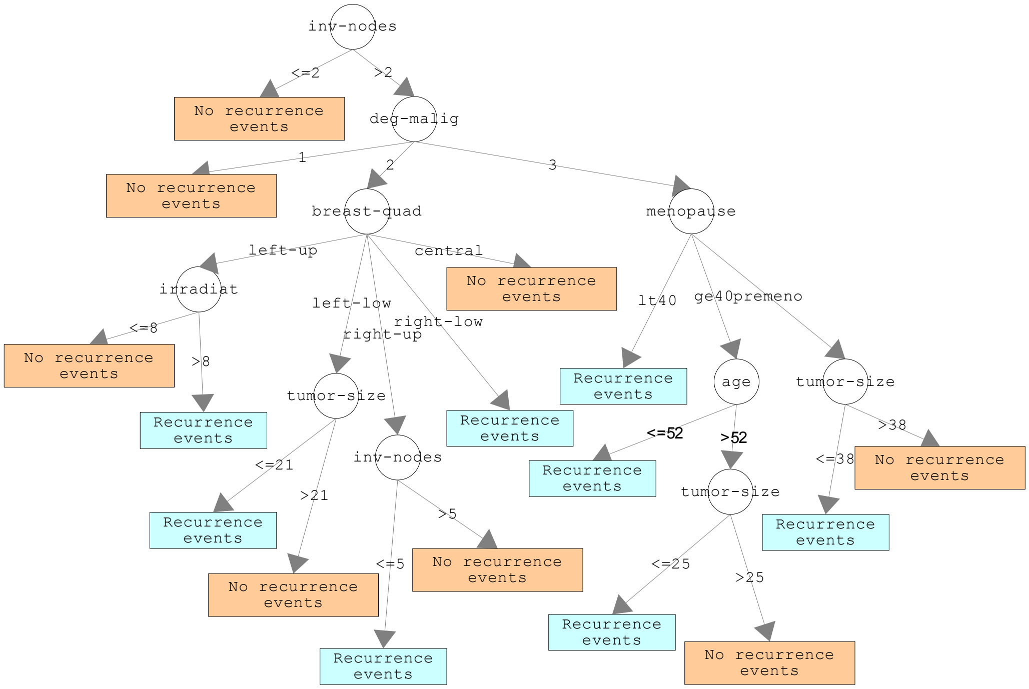

Figure 4. Decision tree induced by PS-C4.5 with δ=0 in the breast cancer

6. References

[1] – Lucas, P. Bayesian Networks in Medicine: a Model-based Approach to Medical Decision Making, [http://citeseer.ist.psu.edu/467626.html].

[2] – Michie, D. On machine intelligence (2^{nd} ed.). Chichester, UK: Ellis Horwood, 1986.

[3] – Newman, D.J., Hettich, S., Blake, C.L., Merz, C.J. (1998). UCI Repository of Machine Learning databases [http://www.ics.uci.edu/~mlearn/MLRepository.html].

[4] – Nuñez, M. The use of background knowledge in decision tree induction. Machine Learning, 6:231-250, 1991.

[5] – Podgorelec, V., Kokol, P., Stiglic, B., Rozman, I. Decision trees: an overview and their use in medicine. J Med Syst. 2002 26(5):445-63.

[6] – Quinlan, J.R. C4.5: Programs for Machine Learning. Morgan Kaufmann, San Mateo, CA., USA, 1993.

[7] – Quinlan, J.R.. Induction of decision trees. Machine Learning, 1(1):81-106, 1986.

[8] – Shannon, C. and Weaver, W. The mathematical theory of communication. University of Illinois Press, Urbana, IL, USA, 1948.

[9] – Shiffman, R.N. Representation of clinical practice guidelines in conventional and augmented decision tables. J Am Med Informatics Assoc 1997; 4:382-93.