# A Theoretical Framework for Decision Trees in Uncertain Domains: Application to Medical Data Sets

B. Crémilleux

GREYC, CNRS - UPRESA 1526
Université de Caen
Esplanade de la Paix
F-14032 Caen Cédex France
cremilleux@info.unicaen.fr

C. Robert

Institut de Recherche
en Mathématiques Appliquées
Université Joseph Fourier
BP 53 X
F-38041 Grenoble Cédex France
crobert@mail-serv.inrialpes.fr

**Abstract.** Experimental evidence shows that many attribute selection criteria involved in the induction of decision trees perform comparably. We set up a theoretical framework that explains this empirical law. It furthermore provides an infinite set of criteria (the C.M. criteria) which contains the most commonly used criteria. We also define C.M. pruning which is suitable in uncertain domains. In such domains, like medicine, some sub-trees which don't lessen the error rate can be relevant to point out some populations of specific interest or to give a representation of a large data file. C.M. pruning allows to keep such sub-trees, even when keeping the sub-trees doesn't increase the classification efficiency. Thus we obtain a consistent framework for both building and pruning decision trees in uncertain domains. We give typical examples in medicine, highlighting routine use of induction in this domain even if the targeted diagnosis cannot be reached for many cases from the findings under investigation.

## 1. Introduction

Decision trees have been used successfully for many different decision making and classifications tasks. Medicine is an important application domain for such methods (see for example [1], [21] and [13]). Broadly speaking, a decision tree is built from a set of training data having attribute values and a class name. The result of the process is represented as a tree in which nodes specify attributes and branches specify attribute values. Leaves of the tree correspond to sets of examples with the same class or to elements in which no more attributes are available. Construction of decision trees is described, among others, by Breiman *et al.* [2] who present an important and well-known monograph on classification trees. A number of standard techniques have been developed in the machine learning community, like the basic algorithms ID3 [32] and CART [2]. A survey of different methods of decision tree classifiers and the various existing issues are presented in Safavian and Landgrebe [36].

In induction of decision trees various attribute selection criteria are used to estimate the quality of attributes in order to select the best one to split on. We will see in Section 2 that considerable research effort has been directed towards comparison of different attribute selection criteria in real world domains ([29] and [5]). It appears that most commonly used criteria perform comparably: this is an empirical law. We cannot escape from setting up a theoretical framework that explains this empirical

law. To achieve this, we write down (in Section 2) the basic constraints of the problem. We derive from them an infinite set of criteria which we call C.M. criteria (concave-maximum or convex-minimum criteria). We will see that the most commonly used criteria which are the Shannon entropy (in the family of ID3 algorithms) and the Gini criterion (in CART algorithm), are C.M. criteria; we can predict at a theoretical level that all C.M. criteria yield similar trees.

In medicine, as in many areas, we are sure a priori that it is impossible to build a tree that correctly classifies all the examples. In such situations, decision tree algorithms tend to divide nodes having few examples and a main drawback appears (see [2], [31], [6] and [37]): the resulting trees tend to be very large and overspecified. Some branches, especially towards the bottom, are due to sample variability and are statistically meaningless (one can also say that they are present due to noise in the sample). Such branches must either not be built or be pruned. If we do not want to build them, we have to set out rules to stop the building of the tree. We know it is better to generate the entire tree and then to prune it (see for example [2] and [16]). In Section 3, we propose a pruning method (called C.M. pruning) suitable in uncertain domains. C.M. pruning builds a new attribute binding the root of a tree with its leaves, the attribute's values corresponding to the branches leading to a leaf. It permits computation of the global quality of a tree. The best sub-tree for pruning is the one that yields the highest quality pruned tree. This pruning method is not tied to the use of the pruned tree as a classifier. Thus we have a consistent framework for both building and pruning decision trees.

In Section 4, we present examples in medical domains where we routinely use decision trees either as a statistical descriptive tool allowing a representation of a large data set, or to point out some populations of specific interest. We compare trees pruned with C.M. pruning to hand-made pruned trees.

## 2. Building Decision Trees

One can find many experiments in the literature that compare several criteria with the intention of giving prominence to the suitable ones (see [26], [14], [27] and [25]). Mingers [29] compares the trees induced by ten selection criteria using data sets corresponding to four fields. These ten criteria all comply either with functions using impurity measures or with $\chi^2$. He notes that the trees constructed with $\chi^2$ seem to be a little more dense, but without being able to give any satisfactory explanation. Further we recall he notes that by choosing at random one attribute for each node, the induced trees hold around twice the amount of nodes as trees produced with other classical criteria. In a more recent paper, Buntine and Niblett [5] present additional results. They conclude that "the entropy criterion is statistically indistinguishable from the Gini criterion" ([5], p. 82).

Let us formulate the question of the attribute selection criterion. Let us consider a node $\Omega$ and let $Y_1,\ldots,Y_p$ be the attributes under study. An attribute selection criterion consists in looking for an extremum of a function $\psi$; let us formulate it as a minimum search problem:

$$\psi(Y_{i_0}) = \min \{ \psi(Y_i) ; i = 1..p \}$$

where $Y_{i_0}$ denotes the attribute selected by the criterion $\psi$.

We present a straightforward sensible set of constraints. Let D be the class that we want to explain, with values $d_1,\ldots,d_k$, and let Y be any attribute defined on the node $\Omega$, with values $y_1,\ldots,y_m$ (see Fig. 1).

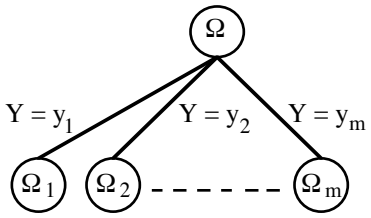


**Fig. 1.** Splitting of a node $\Omega$ using an attribute Y.

i) The minimum value of $\psi$ is reached if and only if the sub-nodes induced by Y are pure with respect to D, that is if and only if two examples with the same value of Y imply that they have the same value of D.
ii) Let P (respectively $P_i$, i = 1…m) be the frequency distribution of D in $\Omega$ (respectively on the set $Y = y_i$, i = 1..m). $\psi$ has its maximum value if and only if $P_1 = \ldots = P_m$ (which implies that $P = P_i$ for all i).
iii) $\psi$ (Y) can be viewed as a combined measure of impurity of the sub-nodes induced by Y. If we want $\psi$ to take into account the respective sizes of the sub-nodes $\Omega_i$ (which is necessary in real world applications), the simplest form for $\psi$ is: $\psi(Y) = \sum_i^m \alpha_i \varphi(\Omega_i)$ where $\Omega_1,\ldots,\Omega_m$ are the sub-nodes yielded by Y, $\alpha_i$ is the rate of $\Omega_i$ in $\Omega$, and $\varphi$ is a function that quantifies the impurity of D in the node $\Omega_i$.

Thus we are led to address the question of defining an impurity measure. A minimal set of constraints is:
(1) The impurity of D in $\Omega$ depends only on the frequency distribution $P = (p_1,\ldots,p_k)$ of D in $\Omega$. Thus a measure of impurity is a function $\varphi$ defined on the set of k-uples with positive coordinates, the sum of which is 1. We will note either $\varphi(P)$ or $\varphi(\Omega)$ the impurity of a node $\Omega$.
(2) $\varphi$ doesn't depend on the way we code D. Thus $\varphi$ has to be equal over all permutations of the components of P. The mathematical term for this is: $\varphi$ is a symmetric function.
(3) $\varphi$ reaches its minimum value $\min\varphi$ if and only if D is a constant function on $\Omega$.
(4) $\varphi$ reaches its maximum value $\max\varphi$ if all values of D are equally frequent.
(5) Combining groups tends to increase the impurity. For example, when we combine r groups $\Omega_i$, D might be constant on each of them, while D might not be constant on the whole group; from i) $\varphi(\Omega_i) = \min\varphi$ and $\varphi(\Omega) \geq \min\varphi$ with $\Omega = \Omega_1 \cup \ldots \cup \Omega_r$.

We have to translate this into a mathematical form and for that, we suggest to refer to current considerations in the fields of statistics (and particularly in ANOVA). We define an intra-group impurity by $\sum \alpha_i \varphi(\Omega_i)$ where $\alpha_i$ is the rate of $\Omega_i$ in $\Omega$; we

suppose $\varphi$ can be linearly split into an intra-group impurity component and an inter-groups impurity component:

$$\varphi(\Omega) = \text{inter-groups impurity} + \sum \alpha_i \varphi(\Omega_i)$$

Since the inter-groups impurity is a positive quantity, $\varphi$ has to satisfy the constraint:

$$\varphi(\Omega) \geq \sum \alpha_i \varphi(\Omega_i)$$

which can be written:

$$\varphi(P) \geq \sum \alpha_i \varphi(P_i)$$

where P (resp. $P_i$) is the frequency distribution of D in $\Omega$ (resp. in $\Omega_i$).

Let us remark that the selection criterion $\psi$ (Y) satisfying iii) is simply the intra-group impurity of the partition of $\Omega$ yielded by Y and that the impurity of $\Omega$ is greater than or equal the average impurity of the sub-nodes. We can also say that this condition states that splitting nodes does not increase impurity.

But since $P = \sum \alpha_i P_i$, the condition $\varphi(P) \geq \sum \alpha_i \varphi(P_i)$ means that $\varphi$ is concave. One can then show [7], that if $\varphi$ is strictly concave (i.e. $\varphi(P) = \sum \alpha_i \varphi(P_i)$ only if $\alpha_i = 1$ for some i, or if $P = P_1 = \ldots = P_m$) and is symmetric, then it satisfies constraints (1) to (5). Finally, any element of the set C of symmetric strictly concave functions is a proper function to define an impurity measure. The most commonly used concave functions and their properties are described for example in Rockafellar [35]. Furthermore, the intra-group impurity $\psi$ defined by iii) satisfies i) and ii).

Finally we have a whole set of possible criteria, which will be called concave minimum criteria. If we had considered a maximum search criterion (replacing $\psi$ by $-\psi$) then we would have had to define purity measures and the proper set would have been the set of symmetric strictly convex functions which would have yielded "convex maximum criteria". We choose to introduce the attribute selection question with impurity, since the notion of impurity is usual in artificial intelligence and it is close to the notion of variance. It is clear that concave minimum criteria and convex minimum criteria are the same, up to a sign -. We will speak of C.M. criteria for both concave-minimum and convex-maximum criteria.

Which C.M. criterion should we choose? The involved impurity (or purity) functions of C.M. criteria have the same concave (or convex) shape and reach their extrema for the same arguments; thus:

1-If $\psi_1$ and $\psi_2$ are two C.M. criteria, Y and Y' are two attributes, then:

$\Delta = [\psi_1(Y) - \psi_1(Y')] \times [\psi_2(Y) - \psi_2(Y')]$ will be positive in most cases and both criteria select the same attribute. Experiments (see Section 4) show that when a criterion $\psi_1$ selects Y, while a second criterion $\psi_2$ selects Y', it appears that $\Delta_1 = [\psi_1(Y) - \psi_1(Y')]$ and $\Delta_2 = [\psi_2(Y) - \psi_2(Y')]$ are small. Thus with a reasonable precision we can consider that $\Delta_1 \approx \Delta_2 \approx 0$ and that the choice between Y and Y' is actually random.

2- We cannot give any theoretical simple condition implying that $\Delta$ is positive.

Let us now consider some commonly used selection criteria.

The most commonly used criterion [32] is that of entropy (which is also called information gain), coming from $\varphi(P) = -\sum p_i \log p_i$ with $P = (p_1,\ldots,p_k)$, which is an impurity measure. It is a C.M. criterion.

The Gini criterion [2] uses $\varphi(P) = 1 - \sum p_i^2$ and is also a C.M. criterion.

The $\chi^2$ criterion (examples are described in [18], [28] and [15]) selects the attribute Y, the chi-square value $\chi^2(D,Y)$ of which is maximum. But $\chi^2(D,Y)$ can be written in the form iii), up to a multiplicative constant N which is the number of elements in the considered node. Thus, using the previous notations,

$$\psi(Y) = \chi^2(D,Y) = \sum \alpha_i \varphi(P_i)$$

where $\varphi(P_i) = N \, ||P_i - P||^2_{(P)}$ Here $||\ldots||_{(P)}$ denotes the metric defined by the diagonal matrix $\begin{pmatrix} 1/p_1 & & 0 \\ & \ddots & \\ 0 & & 1/p_k \end{pmatrix}$.

It appears that $-\varphi$ is a strictly concave function which is not symmetric; $-\varphi(P)$ has its maximum value when $P_1 = \ldots = P_m = P$, and its minimum value when all leaves induced by Y are pure. $-\varphi$ is not an impurity measure and the $\chi^2$ criterion is not a C.M. criterion.

The ratio criterion, deriving from the entropy criterion, is customized to avoid favouring attributes with many values. Actually, in some situations, to select an attribute essentially because it has many values might jeopardize the semantic acceptance of the induced trees ([40] and [24]). The ratio criterion proposed by Quinlan [32] consists in maximizing $\psi(Y) = \dfrac{\varphi(P) - \sum \alpha_i \varphi(P_i)}{\varphi(P_Y)}$ where $\varphi(P)$ is the entropy of P, $\psi$ the associated function and $\varphi(P_Y)$ the entropy of the frequency distribution of Y in the node. It appears that $\psi$ doesn't satisfy condition i) since it can reach its maximum value ($\psi(Y) = 1$) when the sub-nodes yielded by Y are not pure. One example is where D has three values, Y two values and if $Y = y_1$ implies $D = d_1$, while $Y = y_2$ implies $D = d_2$ or $D = d_3$. The ratio criterion is not a C.M. criterion.

Let us note that other selection criteria such as the J-measure [17] are related to other specific issues. The J-measure is the product of two terms that are considered by Goodman and Smyth as the two basic criteria for evaluating a rule: one term is derived from the entropy function and the other measures the simplicity of a rule. Quinlan and Rivest [33] were interested in the minimum description length principle to construct a decision tree minimizing a false classification rate when one looks for general rules and their case's exceptional conditions. This principle has been resumed by Wallace and Patrick [39] who suggest some improvements and show they generally obtain better empirical results than those found by Quinlan. Buntine [4] presents a tree learning algorithm stemmed from Bayesian statistics whose main objective is to provide outstanding predicted class probabilities on the nodes. Kira and Rendell [22] define the algorithm RELIEF for estimating the quality of attributes. The key idea of RELIEF is to assess attributes according to how well their values distinguish among instances that are near to each other. RELIEF is extended by Kononenko [23] to deal with noisy, incomplete, and multi-class data sets. Kononenko

shows that, with some assumptions, the estimates of RELIEF are highly correlated with the Gini criterion. We can also address the question of deciding which sub-nodes have to be built. For a splitting, the GID3* algorithm [12] groups in a single branch the values of an attribute which are estimated meaningless compared to its other values. For building of binary trees, another criterion is twoing [2]. Twoing groups classes into two superclasses so that considered as a two-class problem, the greatest decrease in node impurity is realized. Some properties of twoing are described in Breiman [3]. Always for binary trees, Fayyad and Irani [11] propose the ORT measure. ORT favours attributes that simply separate the different classes without taking into account the number of examples of nodes so that ORT produces trees with small pure (or nearly pure) leaves at their top more often than C.M. criteria. Nevertheless, in uncertain domains, such leaves may be irrelevant and it is difficult to prune them without destroying the tree.

## 3. Pruning Decision Trees

The principal methods for pruning decision trees are examined in [30], [10] and [8]. Most of these pruning methods are based on minimizing a classification error rate when each element of the same node is classified in the most frequent class in this node. These pruning methods are inferred from situations where the built tree will be used as a classifier and they systematically discard a sub-tree which doesn't improve the used classification error rate. We will see that the resulting pruned tree produced by C.M. pruning could be different.

Let us now consider a C.M. criterion. The value of the criterion in a node reflects how appropriately the chosen attribute divides the data. If we consider impurity measures, the smaller the value of $\psi$, the better the split. The value of a criterion permits comparison of divisions of a node, but not of the whole sub-tree built below the node. Fortunately, theoretical considerations embedding C.M. criteria consistently yield a global quality index which will be used for pruning (see [8] for more details). Let us note I (T) the global quality index of the tree T. I (T) measures the difference between the impurity of the root of T and the mean impurity of its leaves, this difference being normalized to a value in [0,1]. I (T) = 1 if and only if all the leaves are pure, and I (T) = 0 if and only if the frequency distributions of D (the class) in the root and in all leaves of T are identical. I (T) doesn't actually depend on which C.M. criterion is used. Moreover if two trees $T_1$ and $T_2$ have the same mean impurity of their leaves but the impurities of their roots are not the same, I ($T_1$) and I ($T_2$) are different. So we can compare two trees built from two different samples of the same population (we will see examples in Section 4).

A straightforward pruning method (that we call C.M. pruning because it goes with using a C.M. criterion to build the tree) perfectly coherent with the building of the tree, consists in pruning the sub-tree T' of T such that T without T' has the highest quality [8]. C.M. pruning produces a family of nested trees spreading from the initial large tree to the tree restricted to its root. We will see in Section 4 that the curve of the global quality index as a function of the number of pruned tree gives a pragmatic method to stop the pruning process. The computational cost of C.M. pruning is particularly low and it is tractable even with large databases.

As previously mentioned, most pruning methods consist of pruning the sub-tree which minimizes a classification error rate. The resulting pruned tree is different from that produced when one uses C.M. pruning. For example C.M. pruning doesn't systematically discard a sub-tree, the classification error rate of which is equal to the rate of the root. In Fig. 2, D is bivalued and in each node the first (resp. second) value indicates the number of examples having the first (resp. second) value of D. This sub-tree doesn't lessen the error rate, which is 10% both in its root or in its leaves; nevertheless the sub-tree is of interest since it points out a specific population with a constant value of D while in the remaining population it's impossible to predict a value for D. The global quality index of this sub-tree is 0.55, which means that it explains 55% of the initial impurity.

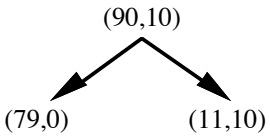


**Fig. 2.** A tree which could be interesting although it doesn't decrease the number of errors.

## 4. Experiments

We have designed induction software called ARBRE, (which means "tree" in French) which produces decision trees using C.M. criteria and which prunes trees with the C.M. pruning method. In this section we describe the results obtained by running ARBRE on two medical data sets. We have seen in the introduction that data sets from the medical domain are frequently used to test induction systems. Uncertainty (i.e. randomness) is basically unavoidable in the medical field. In some domains such as those in the examples below, one knows that the diagnosis is in most cases hardly feasible from the attributes under study. However an appropriately pruned tree may suggest how to separate some sub-populations where the diagnosis is feasible with the involved attributes from the sub-populations where no diagnosis decision can be made without further information. In such situations the tree cannot be used as a classifier; its quality will be poor yet it yields interesting results. Let us remark that in any case pruning is a key point the role of which is to discard the part of the tree that is essentially due to sample variation.

### 4.1. Data Sets

We consider the following two examples:

- *venous thrombo-embolism:* venous thrombosis is a common pathology which can lead to a pulmonary embolism thereby endangering the life of the patient [20]. Among patients having a deep venous thrombosis some will be suffering from pulmonary embolism and others will not. Pulmonary embolism diagnosis rests on more or less complex paraclinical findings and at present it is impossible to accurately predict the risk of this pathology. It is thus impossible to produce any tree which would permit classification of a reasonable part of the examples used to build

it. The main aim of this study is to identify high risk populations to whom complementary examinations would be proposed. Physicians are aware that once these populations, if they exist, are considered aside, it is impossible to predict whether there will or will not be embolism in the remaining population; in other words they are aware that there is no hope of using the tree as a classifier for the whole population of patients with deep venous thrombosis.

This data set TE (thrombo-embolism data set) is composed of all the 1063 patients with deep venous thrombosis treated in the angiology department of the University Hospital at Grenoble (France) and for whom data are reliable. Around fifty percent of these cases were affected by pulmonary embolism (see Table 1). D (the class) is bivalued (embolism versus no-embolism).

• *genetic abnormalities:* genetic abnormalities touch about 1 couple in 600 [19]. Couples with chromosomal segregation allow for two possibilities in their descendants: alternate segregation (the child will be either normal or a healthy carrier) or no-alternate segregation (which implies death or severe handicaps). The genetic cytology department of the Faculty of Medicine in Grenoble has at its disposal the largest European data set on these genetic abnormalities. This data set is constituted of two files:

- the first file G1 is composed of data coming from 86 European Medical Centers and collected from 1975 to 1986.
- the second file G2 is composed of cases published in the medical literature from 1971 to 1984.

**Table 1.** Details of the data files used. "No. of Attributes" indicates the number of attributes including D. "Values / Attributes" are the numbers of values of the attributes; "D" is the number of examples in each class determined by a value of D: in TE (resp. in G1 and G2), the first value is the number of patients who have suffered a pulmonary embolism (resp. alternate segregation) and the second value is the number of patients who never had a pulmonary embolism (resp. no-alternate segregation).

| Data file | No. of Examples | No. of Attributes | Values / Attributes | D |
|---|---|---|---|---|
| TE | 1063 | 7 | 2 - 3 - 4 | 528 - 535 |
| G1 | 2993 | 15 | 2 - 3 | 2646 - 347 |
| G2 | 3247 | 15 | 2 - 3 | 2456 - 791 |

The two files G1 and G2 involve the same attributes. Unlike in the TE set, the frequency distribution of D (alternate segregation versus no-alternate segregation) is highly unsymmetric (see Table 1). The two files were both completely unknown to expert physicians in the domain (such lack of prior knowledge is unusual in medicine but will become more common in the future, due to patient-data management systems). Decision trees provide them with the possibility of acquiring a valuable knowledge of these files. Indeed a decision tree represents all information contained in the data set and oriented by the evolution of the disease towards alternate segregation or no-alternate segregation.

### 4.2. Experimental Procedure

For each data set, we induced both trees using the two C.M. criteria available with ARBRE (entropy and Gini). We performed C.M. pruning with entropy criterion since it is stemmed from the most commonly used impurity measure (see for example [34] and [38]). We pruned each tree until the root was reached. Thus, we obtained a family of nested trees spreading from the initial large tree to the tree restricted to its root. It happens that considering the sequence of quality of these trees allows definition of a pragmatic method to choose a best pruned tree.

We also presented to expert physicians the initial large trees and we asked them to prune these trees; we call this method semantic pruning. We then compared C.M. pruning with semantic pruning.

### 4.3. Results and Discussion

We have predicted in Section 2 that C.M. criteria coming from functions with a common shape and common extrema should to a large extent be interchangeable. This prediction is verified in the three files. For each file, the trees obtained with entropy and with Gini criteria are very similar. For example, the TE tree produced with the entropy criterion is distinguished from the TE tree built with Gini criterion by only two inversions of attributes, but the close values of the criteria for these two attributes explain this phenomenon. Table 2 presents the characteristics of the initial large trees built with entropy criterion.

**Table 2.** Details of the initial large trees built with entropy criterion. The number of nodes includes the leaves.

| | Depth | No. of Nodes | No. of Leaves |
|---|---|---|---|
| TE | 6 | 83 | 59 |
| G1 | 12 | 88 | 48 |
| G2 | 14 | 160 | 86 |

The TE tree exhibits some large leaves with a high embolism risk. These leaves were given confirmation by angiologists though most of them had not been previously brought to light. The trees built from G1 and G2 trees also exhibited relevant leaves with high frequency of alternate segregation. Let us remark that given the large number of attributes and the dissymmetry of the distribution of D, it would have been difficult to use other classical statistical descriptive methods. Finally it happened that though the data in the G1 and G2 files had not been collected in the same way, they yield trees with a close resemblance where the same attributes are selected at the same level of each tree. Here, trees used as descriptive tools yielded the conclusion that the two files could be combined.

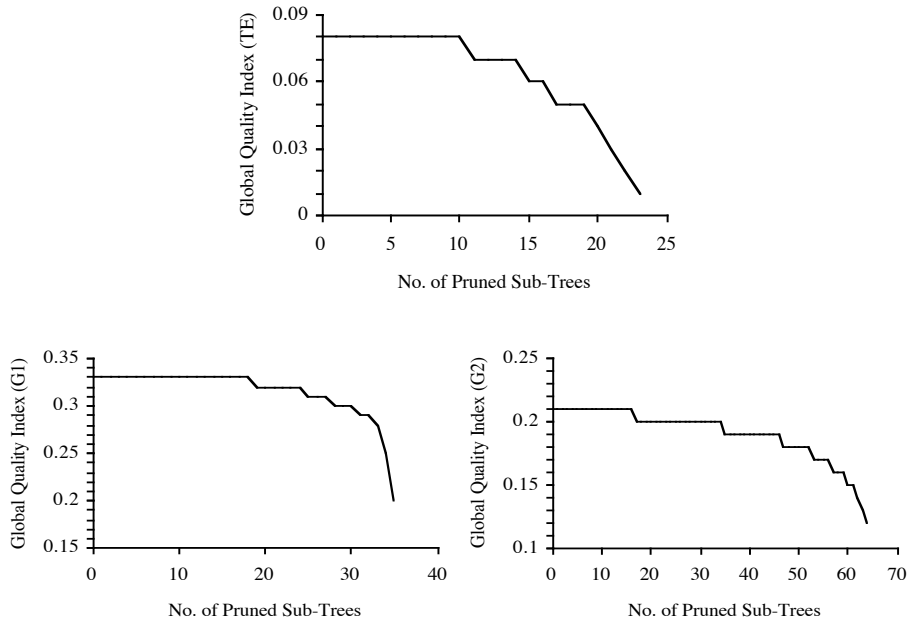


**Fig. 3**. For each tree, advancement of the global quality index according to the number of pruned sub-trees.

Let us now come back to the pruning stage. Fig. 3 is the graphical representation of the global quality index as a function of the number of pruned sub-trees. The global quality index of the TE tree is low (8% of the impurity of the root is explained by the considered attributes). As previously stated, this value is not so surprising given the great difficulty of pulmonary embolism diagnosis.

Let us now consider the global quality index as a function of the number of pruned trees. We see on the graphical representations in Fig. 3 that several sub-trees can be pruned without lessening the global quality index. In the three files considered, the shape of the curves in Fig. 3 indicates that the knowledge which is supplied by the induced trees is essentially at their tops [9]; thus only the highest parts of these trees are reliable. Furthermore, these curves indicate the relevant stages where pruning can be stopped; more precisely since these curves present flat or nearly flat parts, we can stop pruning when the number of pruned sub-trees is a number ending a flat segment of the curve.

The experts for their part pruned the tree until it was around two times smaller than the original tree. Moreover the trees obtained with semantic pruning were very close or identical to one of the trees provided by C.M. pruning. That is: for each family of nested trees, there is always a tree which is very similar (and even identical in the case of G1) to the tree provided by the semantic pruning. More precisely the trees provided by C.M. pruning which are the nearest to the trees given by semantic pruning all have their global quality index reduced by 1%.

## 5. Conclusion

The C.M. attribute selection criteria and the C.M. pruning can be performed consistently within the same theoretical framework. It contains the most commonly used criteria and it explains why these criteria perform comparably. It is a general framework since on the one hand there are no specific conditions on the attributes and on the other hand it doesn't rely on any specific use of the pruned tree. C.M. pruning allows to keep sub-trees with leaves yielding the determination of relevant decision rules, even when keeping the sub-trees doesn't increase the classification efficiency.

The use of a tree as a classifier is highlighted in the artificial intelligence field. In this paper, we stress on some other aspects of the use of decision trees which are, from our point of view, as important as the first, especially with the development of data collection in hospitals. More precisely, a tree built from a data set is an efficient description oriented by an a priori classification of its elements. Pruning the tree discards overspecific information to get a more legible description. A tree can also be built to distinguish some sub-populations of interest in large populations. Here only some leaves of the pruned tree will be considered for further investigation. Let us note that in all the situations described above, pruning is a key point.

Further work has to be done to compare more precisely C.M. pruning with other pruning techniques on consistent data sets. Let us remark that a limit of the global quality index is that its definition doesn't take into account the size of the tree hence the risk associated with the leaves. Another direction is to include such parameters.

## References


[1] Babic, A., Krusinska, E., & Strömberg, J. E. (1992) Extraction of diagnostic rules using recursive partitioning systems: a comparison of two approaches. *Artificial Intelligence in Medicine 4,* 373-387.

[2] Breiman, L., Friedman, J. H., Olshen, R. A., & Stone, C. J. (1984) Classification and regression trees. Wadsworth. Statistics probability series. Belmont.

[3] Breiman, L. (1996) Some properties of splitting criteria (technical note). *Machine Learning 21,* 41-47.

[4] Buntine, W. (1992) Learning classification trees. *Statistics and Computing 2,* 63-73

[5] Buntine, W., & Niblett, T. (1992) A further comparison of splitting rules for decision-tree induction. *Machine Learning 8*, 75-85.

[6] Catlett, J. (1991) Overpruning large decision trees. *In proceedings of the Twelfth International Joint Conference on Artificial Intelligence IJCAI 91.* (pp 764-769). Sydney, Australia.

[7] Crémilleux, B. (1991) Induction automatique : aspects théoriques, le système ARBRE, applications en médecine. Ph D thesis. Joseph Fourier University. Grenoble (France).

[8] Crémilleux, B., & Robert, C. (1996) A Pruning Method for Decision Trees in Uncertain Domains: Applications in Medicine. *In proceedings of the workshop Intelligent Data Analysis in Medicine and Pharmacology, ECAI 96.* (pp 15-20). Budapest, Hungary.

[9] Crémilleux, B., & Zreik, K. (1996) Le rôle de l'interaction personne-système lors de la production d'arbres de décision. *In proceedings of the international Conference on Human-System Learning CAPS 96.* (pp 20-31). Caen, France.

[10] Esposito, F., Malerba, D., & Semeraro, G. (1993) Decision tree pruning as search in the state space. *In proceedings of European Conference on Machine Learning ECML 93.* (pp 165-184). Vienna (Austria), P. B. Brazdil (Ed.). Lecture notes in artificial intelligence. N° 667. Springer-Verlag.

[11] Fayyad, U. M., & Irani, K. B. (1992) The attribute selection problem in decision tree generation. *In proceedings of Tenth National Conference on Artificial Intelligence.* (pp 104-110). Cambridge, MA: AAAI Press/MIT Press.

[12] Fayyad, U. M. (1994) Branching on attribute values in decision tree generation. *In proceedings of Twelfth National Conference on Artificial Intelligence.* (pp 601-606). AAAI Press/MIT Press.

[13] File, P. E., Dugard P. I., & Houston, A. S. (1994) Evaluation of the use of induction in the development of a medical expert system. *Computers and Biomedical Research 27,* 383-395.

[14] Gams, M., & Petkovsek, M. (1988) Learning from examples in the presence of noise. *In proceedings of Eighth International Workshop Expert Systems and Their Applications.* (pp 609-624). Avignon, France.

[15] Gascuel, O., & Caraux, G. (1992) Statistical significance in inductive learning. *In proceedings of the Tenth European Conference on Artificial Intelligence ECAI 92.* (pp 435-439). Vienne, Austria.

[16] Gelfand, S. B., Ravishankar, C. S., & Delp, E. J. (1991) An iterative growing and pruning algorithm for classification tree design. *IEEE Transactions on Pattern Analysis and Machine Intelligence 13(2),* 163-174.

[17] Goodman, R. M. F., & Smyth, P. (1988) Information-theoretic rule induction. *In proceedings of the Eighth European Conference on Artificial Intelligence ECAI 88.* (pp 357-362). München, Germany.

[18] Hart, A. (1984) Experience in the use of an inductive system in knowledge engineering. In M. Bramer (Ed.), *Research and development in expert systems*. Cambridge University Press.

[19] Jalbert, P., Jalbert, H., & Sele, B. (1988) Types of imbalances in human reciprocal translocations: risks at birth. The cytogenetics of mammalian rearrangements, Alan R. Liss. 267-291.

[20] Janssen, F., Schachner, J., Hubbard, J., & Hartman, J. (1987) The risk of deep venous thrombosis: a computerized epidemiologic approach. *Surg. Am.*

[21] Kern, J., Dezelic, G., Dürrigl, T., & Vuletic, S. (1993) Medical decision making based on inductive learning method. *Artificial Intelligence in Medicine 5*, 213-223.

[22] Kira, K., & Rendell, L. (1992) A practical approach to feature selection. *In proceedings of the International Conference on Machine Learning*. (pp 249-256). Aberdeen, D. Sleeman & P. Edwards (Eds). Morgan Kaufmann.

[23] Kononenko, I. (1994) Estimating attributes: analysis and extensions of RELIEF. *In proceedings of European Conference on Machine Learning ECML 94*. (pp 171-182). Catania (Italy), F. Bergadano & L De Raedt (Eds.). Lecture notes in artificial intelligence. N° 784. Springer-Verlag.

[24] Kononenko, I. (1995) On biases in estimating multi-valued attributes. *In proceedings of the Fourteenth International Joint Conference on Artificial Intelligence IJCAI 95*. (pp 1034-1040). Montréal, Canada.

[25] Liu, W. Z., & White, A. P. (1994) The importance of attribute selection measures in decision tree induction. *Machine Learning 15,* 25-41.

[26] Lopez de Mantaras, R. (1991) A distance-based attribute selection measure for decision tree induction. *Machine Learning 6,* 81-92.

[27] Marshall, R. (1986) Partitioning methods for classification and decision making in medicine. *Statistics in Medicine 5*, 517-526.

[28] Mingers, J. (1986) Expert systems - experiments with rule induction. *Journal of the Operational Research Society 37(11),* 1031-1037.

[29] Mingers, J. (1989) An empirical comparison of selection measures for decision-tree induction. Machine Learning 3, 319-342.

[30] Mingers, J. (1989) An empirical comparison of pruning methods for decision-tree induction. *Machine Learning 4,* 227-243.

[31] Niblett, T. (1987) Constructing decision trees in noisy domains. *In proceedings of 2nd European Working Sessions on Learning EWSL 87*. (pp 67-78). Bled (Yugoslavia), Sigma Press. Wilmslow.

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

[33] Quinlan, J. R., & Rivest, R. L. (1989) Inferring decision trees using the minimum description length principle. *Information and Computation 80(3),* 227-248.

[34] Quinlan J. R. (1993) C4.5 Programs for Machine Learning. San Mateo, CA. Morgan Kaufmann.

[35] Rockafellar, R. T. (1970) Convex analysis. Princeton University Press. Princeton. New Jersey.

[36] Safavian, S. R., & Landgrebe, D. (1991) A survey of decision tree classifier methodology. *IEEE Transactions on Systems, Man, and Cybernetics 21(3),* 660-674.

[37] Schaffer, C. (1993) Overfitting avoidance as bias. *Machine Learning 10,* 153-178.

[38] Taylor, C. C., Michie D., & Spiegelhalter, D. J. (1994) Machine learning, neural and statistical classification. Ellis Horwood Series in Artificial Intelligence.

[39] Wallace, C. S., & Patrick, J. D. (1993) Coding decision trees. *Mach. Learn.11,* 7-22.

[40] White, A. P., & Liu, W. Z (1994) Bias in Information-Based Measures in Decision Tree Induction. *Machine Learning 15*, 321-329.