What do we retain from a classification tree?
An experiment in image coding

I.C. Lerman and N. Ghazzali

IRISA, Campus de Beaulieu, 35042 Rennes Cédex, France

Abstract

This work concerns the fundamental problem of the significant reduction of classification trees. Its first contribution is to situate the Gordon approach [3], with respect to ours [7, 8, 9]. The latter is based on the detection of significant nodes in the classification tree. Thus, it was necessary to express very accurately, different notions concerning the searched structure in hierarchical classification. These notions are near but non identical; there are: tree of classifications; valued tree of classifications; hierarchy of subsets and associated tree; valued hierarchy of subsets and associated dendrogram. On the other hand, we have built algorithms which enable the passage from one notion to the another one.

A second original aspect of this paper concerns the definition and the algorithmic construction of a hierarchy of relevant subsets, obtained from the definition of the significant nodes of a classification tree.

The latter notion and consequently, the corresponding algorithmic construction, have been pointed in the framework of using hierarchical classification in image coding (scalar quantization) [10].

Keywords: Hierarchical classification; Parsimonious trees; Significant nodes; Image coding.

Quoi retenir d’un arbre de classification?
Un essai en quantification d’image numérisée

Abstract

Ce travail traite du problème fondamental de la réduction significative des arbres de classification. Son premier apport est de situer l’approche de A.D. Gordon [3] par rapport à la nôtre, plus ancienne [7, 8, 9], qui est fondée sur la détection des noeuds significatifs d’un arbre des classifications. Il a dans ces conditions, été nécessaire de cerner plus précisément et d’élaborer les algorithmes de passage, entre des notions proches, mais non identiques, qui sont – dans la littérature en classification – utilisées plus ou moins indépendamment l’une de l’autre. Il s’agit de: arbre des classifications; arbre indicé ou valu des classifications; hiérarchie des parties et arbre associé; hiérarchie indiquée de parties et dendrogramme associé.

Un deuxième apport original correspond à la définition et à la construction algorithmique d’une hiérarchie de parties pertinentes, à partir de la définition des noeuds significatifs.

Cette notion et la construction algorithmique conséquente, se sont imposées à nous lorsque nous avons voulu utiliser la classification hiérarchique à des fins de quantification scalaire en imagerie numérique noir et blanc [10].

Mots clés: Classification hiérarchique; Arbres condensés; Nœuds significatifs; Quantification en imagerie numérique.
1 General introduction

In this paper, we are concerned with the following topics for which we elaborate our views and add a new contribution:

- the *significant* reduction of a classification tree;
- the reduced tree representation, for optimal readability and clear interpretation of results;
- the selection of *interesting* partitions from the reduced tree.

Indeed, since the end of the 1960’s, the reduction of classification trees has been recognized as an imperious necessity for their interpretation — as soon as the size of the classification set exceeds a few dozens — which rapidly becomes difficult and uncertain; even when the coefficients and criteria selected for the tree formation algorithm are perfectly adapted to the nature of the data. The implicit questions asked by the user, data expert, are the following:

- where does the class begin?;
- does this node correspond to the completion of a class at a particular level of information synthesis?;
- relative to the preceding statement, what are the most *significant* distinctions?;
- How does one retain from the tree of classifications, a few, more or less acute *significant* partitions ?.

In [7], we have proposed a scan in an increasing way, of the different levels of the classification tree on a finite set $E$ with the use of an association criterion $S[\omega (E), \pi (E)]$ between an ordinal information $\omega (E)$ relative to the similarities of the elements of $E$ and a given partition $\pi (E)$ on $E$; being understood that each level in the tree defines a partition. $\omega (E)$ is what we call the *preordonnance* on $E$. Two sequences of numbers are considered:

\begin{align}
\{ S[\omega (E), \pi_i] \mid 1 \leq i \leq I \} \\
\{ \tau_i = S[\omega (E), \pi_i] - S[\omega (E), \pi_{i-1}] \mid 2 \leq i \leq I \},
\end{align}

where $\pi_i$, an $i^{th}$ level partition, is all the less acute that $i$ is large.

Let us recall that the common limiting distribution law of the two dual random variables $S[\omega^*(E), \pi(E)]$ and $S[\omega(E), \pi^*(E)]$ is normal $\mathcal{N}(0, 1)$. In this statement $\omega^*(E)$ [resp. $\pi^*(E)$] is a random preorder (resp. partition) uniformly distributed on the set of all the preordinances (resp. partitions) on $E$, having the same cardinal type as $\omega(E)$ [resp. $\pi(E)$] (for more details, see [8, 9]).

A level (resp. a node) is said to be *significant* if it corresponds to a *local maximum*, along the set of increasing tree levels of $S$ [cf. (1)] {resp. the rate of increase $\tau$ of $S$ [cf. (2)]}. In this context, $S$ is called a *global statistic* and $\tau$ a *local statistic* of the levels.

A partition $\pi_{i_0}$ is all the more significant that $S[\omega(E), \pi_{i_0}]$ represents a local maximum for which its value is distinctively different from the values of $S[\omega(E), \pi_i]$ for $i$ belonging to the interval strictly surrounding $i_0$. Of course, there exists an index $i_1$ for which $S[\omega(E), \pi_{i_1}]$ presents an absolute maximum. It is relative to this $\pi_{i_1}$ partition that J.L. Mollière [14] compares the behaviour of our criterion with the one called CCC (Cubic Clustering Criterion) designed for the classification of clusters within an euclidian space.
The latter comparison concerns large real data. On the other hand, but on artificial data sets G.W. Milligan and M.C. Cooper [13] examined the behaviour of 30 criteria — but not including ours — according to their ability to detect the natural partition present in the data. In this experimental analysis, the influence of the aggregation strategy of the hierarchical clustering method, is taken into account.

However, the expert, a data specialist, is not satisfied with a unique partition and retains several, corresponding to different degrees of accuteness. Furthermore, when the expert is presented with a highly significant partition, he will likely retain almost all classes, but not necessarily all! He might prefer, for one or two of them, to descend on the level directly below and replace the class with its subclasses.

The significant node notion is even more interesting, it corresponds, in experimental practices, to a coherent completion of a subclass [cf. §2]. But then again, there is a definite interest in considering the son nodes which lead to the significant node. Also, the father node of the significant node is just as much interesting.

Although, we have started from the evaluation of an adequation of a same partition produced at a given level in the tree of classifications, we now tend to “forget” this notion of level and keep only a system of pertinent nodes. This approach is very flexible if one desires to retain, from these nodes, a partition for which the size of each class is of a given order (e.g. inferior to a threshold). Indeed, in the literature when it is desired to produce classifications where the number of classes is fixed in size, then — to our knowledge — this size constraint problem has not been addressed. Well, this problem proved to be crucial in the introduction of hierarchical classification for scalar quantization in the field of image coding [10] [cf.§3].

If, up until now, we were proposing a tree representation condensed in the levels where appeared a significant node, we propose here a representation based on a system of pertinent nodes, “forgetting” completely the notion of levels of the tree of classifications which precisely permitted its emergence.

Around the notion of statistically significant cluster, with respect to a probability model of no relation, associating to \( \omega(E) \), a random ordonnance \( \omega^*(E) \), it does exist different approaches in, respectively, different contexts; some of them, depend on a particular hierarchical classification method [12], others are more independant of the particular methodology [1].

Our point of view is somewhat different; we consider that the process of building a classification tree, is different in its logical principle from those of determining a partition, or a particular class, which fit the data. Our method in reducing the classification tree is completely independent from the criteria and the algorithm leading to the tree. It adresses to a given hierarchical classifieraton method to point out the interesting nodes that the method produces in its context.

Gordon’s approach [3] — who at the time ignored our contribution — also consists of retaining a node system but from a hierarchy of subsets, not a tree of classifications. For this, he does not consider — like we do — the significance of a particular node permitting the passage from one partition to a less acute one. What he considers, however, is a partition of the set of nodes obtained from a hierarchical partitioning algorithm. For a given class of this partition, he retains the set of maximal nodes, in the sense of the inclusion of the sets represented by these nodes; hence, for a mutually comparable set of nodes, the node nearest of the root is retained. The partitioning is done taking into account the optimization of a criterion [cf.§2.2].

To understand the difference of nature between our approach and Gordon’s; but also and mostly, for a more general aim of classification, we will present the different formal expressions generally adopted from hierarchical classification and from algorithms for passing through these different structure types [11]:
Figure 1: Tree diagram

- tree of classifications;
- valued tree of classifications or dendrogram;
- valued hierarchy of subsets or tree diagram;
- hierarchy of subsets and associated tree.

For this, we have refered — like it's the quasi general case — to the ascending hierarchical construction of the classification tree. But, this reference is not essential.

Let us note that the tree diagram — of which the structure is equivalent to a valued hierarchy of subsets \( \mathcal{H}(\mathcal{O}), h \) where \( h \) represents the height function — on a given finite set \( \mathcal{O} \), is the graphical representation of a tree for which no branches intersect (cf. figure 1). If one designates by \( \{h_k \mid 1 \leq k \leq l\} \), the set of strictly increasing positive values of \( h \), then one can associate to a tree diagram, a classification tree by considering \( l \) horizontal secants of respective heights: \( h_0 + \epsilon, h_1 + \epsilon, \ldots, h_k + \epsilon, \ldots, h_{l-1} + \epsilon \), where \( h_0 = 0 \) is fixed and \( \epsilon \) is a positive number such that: \( \epsilon < \min \{(h_k - h_{k-1}) \mid 1 \leq k \leq l\} \).

Beneath the \((k + 1)^{th}\) secant, is defined the partition of level \( k \), \( P_k \), which is obtained from the maximal nodes lying under this secant; some of these nodes may correspond to leaves. Thus, from the example of figure 1, the partition defined by horizontal line \( h_2 + \epsilon \), is: \( \{\{a, b, c\}, \{d\}, \{e, f\}, \{g, h, i, j\}\} \); the maximal nodes being \( \{d\} \), \( B \), \( D \) and \( E \). Each maximal node covers a subset of \( \mathcal{O} \) and these different subsets form a partition of \( \mathcal{O} \).

If to a classification tree [cf. figure 2] corresponds one and only one hierarchy of subsets \( \mathcal{H}(\mathcal{O}) \). Conversely, to a hierarchy of subsets corresponds what we call a tree of subsets. A sequence of nodes situated on an unique path of this tree, represents a sequence of subsets of \( \mathcal{O} \) [elements of \( \mathcal{H}(\mathcal{O}) \)] strictly monotone with respect to the inclusion relation between subsets. Like for the case of the classification tree, the root of this tree represents \( \mathcal{O} \) and the leaves, the elements of \( \mathcal{O} \). But, the length (in terms of the number of edges) of a complete path starting from the root and ending with a leaf (a terminal node), is not invariant. Let \( \mathcal{A}_{\mathcal{H}}(\mathcal{O}) \) designate the tree of subsets associated to \( \mathcal{H}(\mathcal{O}) \), and, before proposing an algorithm for passing between \( \mathcal{H}(\mathcal{O}) \) and \( \mathcal{A}_{\mathcal{H}}(\mathcal{O}) \), let us give an example.
Consider the following hierarchy of subsets on the set $\mathcal{O} = \{a, b, c, d, e, f, g, h, i, j\}$ of ten elements:

\[
\mathcal{H}(\mathcal{O}) = \{H_0 = \mathcal{O}, H_1 = \{d, e, f, g, h, i, j\}, H_2 = \{d, e, f\}, H_3 = \{g, h, i, j\}, H_4 = \{e, f\}, H_5 = \{h, i\}, H_6 = \{a, b, c\}, H_7 = \{a\}, H_8 = \{b\}, H_9 = \{c\}, H_{10} = \{d\}, H_{11} = \{e\}, H_{12} = \{f\}, H_{13} = \{g\}, H_{14} = \{h\}, H_{15} = \{i\}, H_{16} = \{j\}\}.
\]

(3)

The associated tree of subsets $\mathcal{A}_{\mathcal{H}}(\mathcal{O})$ is represented in figure 3, where the leaves of the tree are directly labeled with the elements of $\mathcal{O}$.

In short, let $T = \mathcal{A}_{\mathcal{H}}(\mathcal{O})$ be the tree of subsets associated to $\mathcal{H}(\mathcal{O})$ and let $a$ be the generic node of level $n_T(a)$. Let $\text{card}_T(a)$ be the cardinal of the subset of $\mathcal{O}$, element of the hierarchy $\mathcal{H}(\mathcal{O})$, represented by the node $a$. The immediate predecessor in $T$ of the node $a$ is denoted by $\text{pred}_T(a)$ with, of course, $\text{card}_T[\text{pred}_T(a)] \geq \text{card}_T(a)$. Finally, using the notation of [5], $(a.1, a.2, \ldots, a.i, \ldots, a.q)$ represents the $q$ immediate descendants of $a$ in an ascending order, that is to say: $\text{card}_T(a.1) \geq \text{card}_T(a.2) \geq \cdots \geq \text{card}_T(a.i) \geq \cdots \geq \text{card}_T(a.q)$.

Then, the recursive algorithm for constructing the tree $T$ has the following steps$^1$:

Etape 0 : on part de la racine $H_0$ ;

faire jusqu'à arrêt ;

si tous les éléments de $\mathcal{H}(\mathcal{O})$ sont visités, alors arrêt ;

sinon, passer à l'étape 1.

Etape 1: traiter le noeud courant $a$ ;

si $\text{card}_T(a) = 1$, auquel cas $a$ est une feuille, alors passer à l'étape 2 ;

sinon, c'est que $\text{card}_T(a) > 1$, auquel cas traiter le descendant immédiat de $a$: prendre le descendant $a.i$ de $a$ dont le cardinal est supérieur à tous les cardinaux de ses $(q - 1)$ frères $a.j$ avec $j = 1, q$ et $j \neq i$, et le mettre dans $a$; puis retourner à l'étape 1.

$^1$ je compte réécrire l'algorithme
We have also expressed in section 1 the general principle of this approach where one starts by determining a partition of the set of nodes for which each of them represents an element of the hierarchy of subsets $H(O)$.

Two methods are proposed for transforming a dendogram into a parsimonious tree. The first — called global — consists of giving an absolute value to the notion of height, thus granting a meaning to the comparison between $h(C)$ and $h(D)$, when $C$ and $D$, elements of $H(O)$, are two disjoined subsets. Then one considers, on the tree diagram (cf. figure 1, §1), $(g - 1)$ horizontal lines, where the height of a given line is strictly within two consecutive values of $h$ on $H(O)$. Thus, one obtains a partition in $g$ classes of the set of nodes, the $e^{th}$ being determined by the subset of nodes lying between the $(e - 1)^{th}$ and the $e^{th}$ horizontal line. From this $e^{th}$ class, $1 \leq e \leq g$, one retains — like mentioned in section 1 — the maximal nodes nearest the root; they are called favored nodes. The resulting reduction of the valued hierarchy of subsets corresponds to replacing the height function $h$ by a height function $p$ including the smallest and the largest values of $h$. One can interpret the input of the functions $h$ and the $p$ on the set $P_2(O)$ of distinct object pairs of $O$; in this case $h$ and $p$ define two ultrametric distances on $O$ and then A.D. Gordon considers a misfit function of the form:

$$S(h, p, g) = \sum_{1 \leq i < j \leq n} w_{ij} \sigma_{ij}(h, p, g),$$  \hspace{1cm} (4)

where $\{w_{ij} \mid 1 \leq i < j \leq n\}$ is a weight function on $P_2(O)$ to be discussed and where $\sigma_{ij}(h, p, g)$ is the resulting loss from replacing $h_{ij}$ with $p_{ij}$.

The choice of an appropriate function $\sigma_{ij}(h, p, g)$ depends on the properties of the tree $(h_{i,j})$: three possibly relevant functions are

$$\sigma_{ij}^S(h, p, g) = (h_{ij} - p_{ij}), \quad \text{where} \quad p_{ij} \leq h_{ij} \quad 1 \leq i < j \leq n. \hspace{1cm} (5)$$

$$\sigma_{ij}^C(h, p, g) = (p_{ij} - h_{ij}), \quad \text{where} \quad h_{ij} \leq p_{ij} \quad 1 \leq i < j \leq n. \hspace{1cm} (6)$$

$$\sigma_{ij}^A(h, p, g) = (p_{ij} - h_{ij})^2, \hspace{1cm} (7)$$

The additive nature of criterion $S(h, p, g)$ [cf.(4)] on increasing node heights, allows to consider, for a fixed $g$ and in the manner of W.D. Fisher [2], dynamic programming for finding the overall optimal solution to the problem. However, the choice of $g$ remains open to discussion. It can be pointed out here, that dynamic programming has been used in a number of qualitative data analysis fields [6][8], chap. 9], etc...

As previously the favored nodes are the maximal nodes in the sense of the order relation on $H(O)$; that is to say, once again, the nearest to the root. The second partitioning process — called local — consists of removing edges from the tree. Noting that a remove ??? edge allows to obtain a descending subtree, the root of which determining precisely a favored node, one only need to remove $(g - 1)$ edges for obtaining $g$ unconnected subgraphs, corresponding to the $g$ groups of internal nodes.

In these conditions, it is not possible to use dynamic programming to minimize the value of the misfit criterion. For a fixed $g$, a good heuristic is needed. Also, the number $g$ of groups is still an open question.

### 2.2 Significant levels and nodes

Let $E$ denote the set of the entities to be classified. Relative to a classical data table crossing a set $O$ of objects with a set $V$ of descriptive variables, $E$ can either represent the object set $O$ or the variable set $V$. 
Étape 2 : traiter le prédécesseur ;
enlever la partie singleton correspondante à la feuille a, de \( \text{pred}_T(a) \) et dans \( a \), mettre \( \text{pred}_T(a) \); retourner à l'étape 1.

Arrêt.

Il est clair que à un arbre de classifications (cf. figure 2), il correspond une et seule hiérarchie de sous-ensembles \( \mathcal{H}(\mathcal{O}) \) et, conséquemment, un arbre de sous-ensembles \( \mathcal{A}_\mathcal{H}(\mathcal{O}) \). Mais contrairement à une hiérarchie de sous-ensembles \( \mathcal{H}(\mathcal{O}) \) il peut correspondre plusieurs arbre de classifications compatible avec la hiérarchie. Ces arbres peuvent être obtenus de manières récursives et numérisées de manière conséquente [11].

En section 2, Gordon’s approach and ours are presented with new graphical and algorithmic developments. In section 3, we will describe an application in scalar quantization for image coding. This application has been a direct stimulation for this research. Section 4 will conclude with future prospects.

2 Tree reduction approaches

In this section, we will develop more precisely and technically the different aspects presented in as it was the general introduction (cf. §1). First, in subsection 2.1, we will summarize Gordon’s approach [3]. In subsection 2.2, we will express our approach in its state of development [7, 8, 9] as it was immediately before the original contribution leading to this article. In subsection 2.3, we will present the new algorithmic idea, based on the detection of significant nodes, for representing a condensed tree of subsets stemmed from the tree of classifications. This idea, which has been implemented, is illustrated in an example and is used for the application in scalar quantization.

2.1 Gordon’s approach

The basic structure considered is a valued hierarchy of subsets on a finite set \( \mathcal{O} \) of objects. We have seen that, to a such structure, it can be associated a valued tree of classifications.
\( F = P_2(E) \) will indicate the set of pairs or two element subsets of \( E \). Then suppose that on \( E \), is already established a notion of similarity \( S \). The latter corresponds in almost cases of data classification, to a numerical valuation of \( F \). If the datum is a dissimilarity index \( D \) on \( E \), then naturally we set \( D = -S \).

The information of a partition \( \pi \) on \( E \) can be interpreted at the \( F \) level, in terms of a total preorder, with two classes denoted by \( R(\pi) \) and \( S(\pi) \). \( R(\pi) \) [resp. \( S(\pi) \)] is the set of joined (resp. disjoined) pairs.

A general principle of our approach is to reduce the problem to a comparison of two identical combinatorial structures. For this aim, we consider the preorder on \( E \), associated to the similarity \( S \). It is a total preorder on \( F \), defined as follows:

\[
\forall (p, q) \in F \times F \), \( p \leq q \Leftrightarrow S(p) \leq S(q).
\]

Let \( \omega(E) \) be the induced ordinal structure. Then, two cases are considered, depending on whether \( \omega(E) \) is a total order or a preorder on \( F \). The second case is where the similarity function \( S \) is not an injection in its definition on \( F \).

Following what it has been expressed in section 1, the general expression of the global statistic \( S[\omega(E), \pi(E)] \) is:

\[
S[\omega(E), \pi(E)] = \frac{s[\omega(E), \pi(E)] - \mathcal{E}\{s[\omega(E), \pi^*(E)]\}}{\sqrt{\text{var}\{s[\omega(E), \pi^*(E)]\}}}
\]

where \( s[\omega(E), \pi(E)] \) is a raw association coefficient which matches the set theoretic representation of \( \omega(E) \) and \( \pi(E) \), at the level of the cartesian product of \( F \times F \), and where \( \pi^*(E) \) is a random partition on \( E \) associated to \( \pi(E) \) under cardinality conditions (for more details, see [7, 8, 9]). On the other hand, \( \mathcal{E}(\cdot) \) and \( \text{var}(\cdot) \) denote respectively, the mean and the variance of the random raw index.

Relative to expression (1) of section 1, let us denote:

\[
S_i = S[\omega(E), \pi_i]
\]

and let the numerical function on the set \( \{1, 2, \cdots, i, \cdots, J\} \) of the tree levels be:

\[
i \mapsto S_i
\]

Such a function has a global tendency towards first increasing up to with its maximum value with a slope near the horizontal. After a small stage, it suddenly decreases. Every thing is as if, in the succession of aggregations, one sought to attain the best classification: the one which, in a small number of classes, constitute the most general summary. From the latter, any further aggregation can only be unnatural and is thus penalized by a great decrease of \( S \).

The situation just described about the behaviour of \( S \), where only one dominant mode is present in the distribution \( \{S_i \mid 1 \leq i \leq I\} \), is the most frequent. But, we have also observed some cases where two or three consistent modes coexist and indicate strong states of equilibrium in the automatic synthesis.

It should also be noted that within a same interval — of the increasing succession of levels — of global growth of \( S \), certain unions of classes can be accompagnied by a local decrease of \( S \). This expresses that, considered globally, the obtained partition after such a union, is less in agreement with the preorder \( \omega(E) \), than the preceding one which will be retained if a choice has to be made between the two.

If the behaviour of the global statistic allows us to recognize the few most significant levels, the behaviour of the local statistic \( \tau \) [cf.(2) §1] allows to detect the significant nodes of
the classification tree. Let us consider the numerical function on the set \( \{2, \cdots, i, \cdots, I\} \) of the
tree levels:

\[
i \mapsto \tau_i = S_i - S_{i-1}
\]

(12)

\( \tau_i \) represents the criterion's growth rate for the passage between partition \( \pi_{i-1} \) of level
\((i - 1) \) and partition \( \pi_i \) of level \( i \).

A very valuable experience has shown that the distribution \( \{\tau_i \mid 2 \leq i \leq I\} \) of \( \tau \), along
the succession of levels is such that, its value increases when a class in formation is confirmed
semantically, and decreases appreciably just before the end of a consistent class constitution,
for the benefit of the birth or the embryonic growth of another class. The levels associated
to local maxima of the distribution \( \{\tau_i \mid 2 \leq i \leq I\} \), therefore correspond to class completion
levels. We will thus retain as most significant, the nodes defined by these levels. And, up until
now, it is the tree reduced to its levels that we usually represent graphically. The significant
nodes are — in this representation — marked by an asterisk accompanying the level number
where they appear; other nodes are simply marked by their level number.

The local minima of \( \{\tau_i \mid 2 \leq i \leq I\} \) are also revealing for the data expert. They express
at different levels of synthesis the relative oppositions between the subclasses of a larger class.
The interpretation of a local minimum of \( \tau \), requires an all the more acute comprehension of
the data, that the level \( i \) is higher.

Thus, our method for the recognition of significant levels and nodes of a tree of class-
ifications does not make any reference to an outside model. On the contrary, the approach
of P.H.A. Sneath [17], consisting in the proposition of separation tests between two classes,
requires parametric models, exogenous to the data, for which relevance must be justified.

Associations being judged relatively from inside the data, our method gives a dynamic
interpretation; that is to say, following step by step the class formations of the tree of classi-
fications. Moreover, if the magnitude of the values attained by \( S_i \) (resp. \( \tau_i \)) has a certain interest;
it should be known that what really matters, is the evolution of the values \( \{S_i \mid 1 \leq i \leq I\} \)
(resp. \( \{\tau_i \mid 2 \leq i \leq I\} \)).

This concern to recognize pertinent and well identified components in the construction
of a tree of classifications, does exist in other approaches of purely geometrical inspiration
[e.g.,[4, 13]]. However, these, use the same criteria that prevailed for class formations. Without
denying anything about the interest of these techniques, we believe that there exist good criteria
for class emergence and good criteria for their evaluation, but, that they are not necessarily the
same!

Also and above all, the evaluation criterion must have a very general character; that is, to
be completely independent of the class formation criterion. It is clearly the case for our criterion
\( S[\omega(E), \pi] \).

Finally, the evaluation criterion must correspond to a clear formal conception and to a
statistical significance basis. This is also the case for our criterion.

2.3 Tree of pertinent subsets

The search for pertinent subsets based on the notion of significant nodes, in accordance to the
intuitive introduction [cf. §1], in the classification tree is done by only considering the branches
that originate from or end to a significant node. The formation of node \( a \), as illustrated in
figure 4, is retained only if it is itself significant or an immediat descendant of a significant
node [cf. figure 4(1)] or a father of a significant node [cf. figure 4(2)]. Otherwise [cf. figure
4(3)], the formation of a node \( a \) is not taken into account. A significant node is represented by
with an asterisk in front of its number.
This process is repeated, in a recursive manner, for all tree nodes. Hence, we obtain a succession of subsets which are the most interesting in the tree. This succession of subsets will be represented in the form of a tree, such that all the paths from the root to the leaves have the same length. The total depth is equal to the number of obtained partitions, and for which, each level, called physical, corresponds to one of these partitions. The first level corresponds to the finest partition which contains the greatest number of classes. As for the last, which is just before the root of the tree of subsets, it corresponds to the most rough partition with more than one class, having the smallest number of classes. Such representation of the obtained partitions seems to us the most reduced, for a relevant expression of the respective compositions of the different classes obtained in an agglomerative procedure. A given level having only a physical meaning, it results a great flexibility to extract several partitions from this structure.

In the following, we first describe how the physical levels of the tree are formed and then, we will present the algorithm for generating the partitions. Figure 5 represents the classification tree, stemmed from an ascending hierarchical classification method, on a set of 42 objects containing 12 significant nodes preceded by an asterisk.

Our starting point is the list of nodes of a classification tree where those that are significant are already specified; that is a prefix polish notation of the tree with the list of the significant nodes. The leaves (elements to be classified) are at level 0 and the tree is processed from the root to the leaves.

If a particular tree node $a$ verifies conditions (1) or (2) [cf. figure 4], its physical level is increased by one. On the contrary, when condition (3) [cf. figure 4] is true, its physical level is set to either the highest value of its two descendants $a.1$ and $a.2$, or to 1 if $a$ is a leaf. Hence, by recursion on the nodes, the sought after subsets are obtained. This process is described by function traititns which detailed below. Figure 6 represents the tree of pertinent subsets resulting from the reduction of the one, illustrated in figure 5. Nine partitions are obtained. The first and most acute is $P_1 = \{\{1 - 7\}, \{8 - 12\}, \{13 - 14\}, \{15 - 16\}, \{17\}, \{18\}, \{19\}, \{20 - 22\}, \{23 - 24\}, \{25\}, \{26\}, \{27 - 29\}, \{30 - 31\}, \{32 - 34\}, \{35\}, \{36 - 42\}\}$, formed of 16 classes. The last partition, before the root, is $P_9 = \{\{1 - 22\}, \{23 - 42\}\}$ formed of 2 classes. An intermediate partition is, for example, $P_7 = \{\{1 - 7\}, \{8 - 12\}, \{13 - 22\}, \{23 - 42\}\}$ formed of 4 classes.

The algorithm for generating these partitions is now described:

reducits algorithm
/* proceeds with the tree reduction by determining the best partitions based on significant nodes */
BEGIN
    root = pol2node(); /* builds tree from the polish notation */
    traititns(root); /* reduces the tree starting from the root */
node2pol(root); /* converts the tree to polish notation */
END

Function \textit{traite\.ns}(a) is for processing a node \(a\) within a tree \(T\):

\textbf{traite\.ns}(a)
BEGIN
\hspace{1em}IF \(\text{card}_T(a) = 1\) THEN
\hspace{2em}\(n_T(a) = 0\);
\hspace{2em}return;
\hspace{1em}ELSE
\hspace{2em}traite\.ns(a.1);
\hspace{2em}traite\.ns(a.2);
\hspace{2em}IF \(a\) verify condition (1) or (2) THEN
\hspace{3em}\(n_T(a) = \max[n_T(a.1), n_T(a.2)] + 1\);
\hspace{2em}ELSE
\hspace{3em}\(n_T(a) = \max(\max(n_T(a.1), n_T(a.2)), 1)\);
\hspace{1em}ENDIF
ENDIF
END

Another problem concerns the class sizes of partitions stemmed from the classification tree. It can fluctuate considerably and, thus, can affect the quality of the classification results depending on the type of data. For a given classification tree, we propose an algorithm called \textit{reduc\.lg} that proceeds to form pertinent subsets while imposing a threshold on the number of elements composing each class (tree node). Only the \textit{traite\.lg} function will be described, which implements this constraint. The algorithm starts with the prefixed polish notation of a classification tree \(T\) for which we want to extract a partition with \(k\) classes having each \(s\) elements or less. The result produced by the algorithm is a file containing a description of \(h\) classes \((h \geq k)\) of this partition which is function of the chosen threshold \(s\).

For a given node \(a\) of tree \(T\), function \textit{traite\.lg}(a, \(s\)) follows:

\textbf{traite\.lg}(a, \(s\))
BEGIN
\hspace{1em}IF \(\text{card}_T(a) \leq s\) THEN
\hspace{2em}accept the class formed by the leaves of subtree \(a/T\)
\hspace{1em}ELSE
\hspace{2em}TRAITE\.LG(a.1, \(s\));
\hspace{2em}TRAITE\.LG(a.2, \(s\));
\hspace{1em}ENDIF
END

\textit{reduc\.ns} and \textit{reduc\.lg} can be used separately or combined to form algorithm \textit{reduc\.nslg} now described briefly:

\textbf{reduc\.nslg} algorithm
BEGIN
\hspace{1em}root = pol2node();
\hspace{1em}traite\.ns(root);
\hspace{1em}node2pol(root);
\hspace{1em}traite\.lg(root, \(s\));
END
These algorithms can be applied to any classification tree stemmed from an ascending hierarchical construction algorithm (AHC).

For this work, we have processed classification trees obtained from an ascending family of hierarchical classification methods (LLA) $\theta \leq 1$ based on the likelihood of linkages method [10] in the context of scalar quantization for image coding [15].

3 Application in scalar quantization

The problem of data compression for image transmission and storage has proved to be of major interest in the field of quantization for image coding. Several techniques have been developed to answer this need and allow great compression of data — which are in most cases very voluminous — with good subjective quality in the perception of reconstructed images. These methods of scalar and vector quantization [16] rely on classification methods and offer the advantage of simple decoding.

The object of our work is to unite within the same conceptual framework, quantization [15] and classification [10] by presenting for the latter a panorama on several methods of hierarchical and non-hierarchical classification for which we show all their importance and richness.

Indeed, our interest lies on fixed images of 256 grey levels and with a resolution of 8 pixels/mm, that we wish to compress using one or more classification methods with the double objective of good image reconstruction and interesting data compression.

Recall that an image can be assimilated to an array $T(x,y)$ where $x$ and $y$ are the coordinates of the array elements called pixels (picture elements). To this array, is assigned a function $L(x,y)$ called the luminosity such that

$$0 \leq L(x,y) \leq 255 = 2^8 - 1,$$

where black is represented by 0 and white, by 255. These limits are generally fixed independently of the image size which is more easily modifiable. We work with discrete images of sizes $512 \times 512$ and $576 \times 720$.

The statistical representation of an image is then a one-dimensional histogram

$$\{(l, n_l) \mid 0 \leq l \leq 255\}$$

(14)

where $n_l$ is the number of pixels which have luminosity $l$ and, depending on the image size:

$$\sum_{l=0}^{255} n_l = N = 512 \times 512 \text{ or } 576 \times 720$$

Perfect reconstruction is attained if every pixel is coded with 8 bits. Thus, the problem consists of limiting the number of bits (less than 8) while reconstructing the original image with the best possible visual perception.

We apply on the histogram formulated in (14), thresholding techniques resulting from different automatic classification methods to obtain a judicious subdivision of the scale defined in (13) into $k$ intervals such that: $\sigma_k = (l_0, \ldots, l_{j-1}, l_j, \ldots, l_{k-1}, l_k = 255)$ and that $I_j = [l_{j-1}, l_j[. By choosing for each $I_j$, $1 \leq j \leq k$, a relevant representing point by means of an objective criterion, scalar quantization of an image is realized and constitute a fundamental step towards vector quantization.

For some classification algorithms, we have been faced to the problem of unconnected classes. Another problem that surfaced for all algorithms concerns the high amplitude variation of classes that correspond here to intervals. Because formed classes have almost equal weights (number of associated pixels), the amplitude variation (number of associated luminosity levels)
can change considerably from one class to another one and, then, leads to a deterioration in the perception quality of the restored images. Hence, to solve these types of problems, we propose an approach allowing the creation of a tree of pertinent subsets [cf. §2.3] from the classification tree.

The example that we present next is a fixed image, called voiture (supplied by the C.C.E.T.T.\textsuperscript{2}), in 256 grey levels and of size 512 \times 512. On the histogram of this image, we have applied an element of the family (LLA), \( \epsilon = 0.5 \), corresponding to an optimum, based on a paired comparison procedure described by the C.C.I.R.\textsuperscript{3}, of image quality for the compressed image in 16 grey levels illustrated in figure 7. We insist on the fact that the object of this work is not to attain phenomenal data compression, which is more the objective of vector quantization. The obtained partition \( P = \{ \{0-27\}, \{28-59\}, \{60-73\}, \{74-85\}, \{86-94\}, \{95-102\}, \{103-112\}, \{113-122\}, \{123-133\}, \{134-147\}, \{148-165\}, \{166-182\}, \{183-198\}, \{199-213\}, \{214-235\}, \{236-255\} \} \) formed of 16 classes was used to reconstruct the image of figure 7.

4 Conclusion and future prospects

The criteria and algorithms proposed in this paper, taking into account both their elaboration and their statistical foundations, have the greatest degree of generality. That is to say, they are not dependent on a particular data representation, nor on the indices or criteria used, nor on the algorithms for the construction of the classification tree.

Mostly, we assume an ascending hierarchical building for the tree of classifications. But this is not an obligation! In fact, for a descending construction by successive segmentations, the tree levels will be scanned, not from the leaves to the root; but in a descending manner, from the root to the leaves.

What is significantly new in this work lies, on the one hand, in the comparison of A.D. Gordon's approach relative to ours which is older and, on the second hand, following the elimination of the absolute value of the level notion, with respect to the construction step by step of a classification tree, the passage from a condensed tree of classifications to a tree of pertinent subsets. The latter being always founded on the detection of significant nodes in the classification tree.

By considering this tree of pertinent subsets, one can imagine to determine interesting partitions by retaining a subset of its nodes. The interesting character of a partition will be judged from the criterion \( S[\omega(E), \pi(E)] \). There is, in this approach, a new algorithmic concept to invent.

Criterion \( S[\omega(E), \pi(E)] \) has statistical significance. Indeed, \( S[\omega(E), \pi(E)] \) has very small values for the first tree levels, when there is, so to speak, no inversion between the preordonnance and the one, in two classes, that stemmed from the partition. In other words, the concerned partitions have very little consistency. If one desires a coefficient which can show the similarity of two preordonnance independently of their respective consistency, one can consider

\[
\frac{S[\omega(E), \pi(E)]}{S(\pi(E), \pi(E))}
\]

where \( S(\pi(E), \pi(E)) \) is the criterion value between the following preordonnance \( \pi(E) \) and itself. \( \pi(E) \) has exactly two classes \( S(\pi) \) and \( R(\pi) \) \( [S(\pi) < R(\pi)] \), where \( S(\pi) \) [resp. \( R(\pi) \)] is the

\textsuperscript{2}Centre Commun d'Études de Télédiffusion et Télécommunications de Rennes — France

\textsuperscript{3}Comité Consultatif International des Radiocommunications subject 11/4 doc.156,1986-1990
Figure 5:

Figure 6:

Figure 7:
subset of object pairs separated [resp. joined] by the partition $\pi$.

Finally, one can say that the ideas presented in this paper, which have been latent for the most part, have found precipitation thanks to the constraints imposed by scalar quantization in image coding [cf. §3], from hierarchical classification. The remaining question is now how to apply these ideas in the context of vector quantization for image coding.
References


