Hierarchical Classification for Seabed Characterization

Christophe Osswald and Arnaud Martin

ENSIETA - E^3I^2 - EA3876
2 rue François Verny,
29806 BREST Cedex 09, France
(e-mail: `Christophe.Osswald@ensieta.fr, Arnaud.Martin@ensieta.fr`)

Abstract. The automatic seabed characterization is a difficult problem. Most automatic characterization approaches are based on texture analysis. Indeed, the sonar seabed images present many homogeneous areas of sediment that can be interpretated as a sonar texture.

Here, we optimize the agglomerative hierarchical clustering algorithm to produce homogenous clusters of sediments images, combining known and unknow data.
Keywords: Classification, Sonar, Seabed Characterization.

1 Introduction

The problem of automatic seabed characterization is very important and difficult. The seabed characterization is important in order to make seabed maps for sedimentologists, for autonomous underwater vehicle navigation or pollution. One approach in order to characterize the seabed is the use of a sonar. The main issues with sonar images is those are particularly difficult to characterize by automatic process. The expert has never the certainty to differentiate well the sand from the silt for example: the difference between these sediments comes only from the granulometry that varies continuously.

We first expose the principle of agglomerative hierarchical classification. In section 2 we present the sonar images data and the considered texture analysis. We study the usual clustering methods applied on sonar small-images. Then in section 5 we define a hierarchy quality in order to choose better agregation functions for hierarchical classification. In section 6, we present results of the combination of known and unknow data in order to characterize the sediment of the sonar images.

2 Agglomerative Hierarchical Classification

Agglomerative hierarchical classification (AHC) is a common approach to build a clustering system from a dataset. The algorithm considers the objects of the dataset as trivial clusters of size 1: $d(\{x\},\{y\}) = d(x,y)$. Then, at each step, the algorithm merges the two nearest clusters into a new cluster, and computes the distance between the new cluster and the other ones. The index associated to the cluster $C = A \cup B$ is the dissimilarity $d(A, B)$.

The dissimilarity induced by an indexed hierarchy (*i.e.* dissimilarity between x and y is the smallest index of a cluster containing x and y) is an *ultrametric*.

The natural clustering system of a dissimilarity, in the way of Jardine and Sibson [Jardine and Sibson, 1971], is composed of the maximal cliques of its threshold graphs, indexed by the diameter of the clusters. So a set A is a cluster for the dissimilarity d with an index λ if:

(i) there exists x and y in A such that $d(x, y) = \lambda$,
(ii) u and v in A brings $d(u, v) \leqslant \lambda$,
(iii) for any z not in A there exists t in A such that $d(z, t) > \lambda$.

The indexed clustering system induced by an ultrametric is an indexed hierarchy. It is well-known that ultrametrics and indexed hierarchies are in bijection.

Let d be a dissimilarity on X and used as a dissimilarity on the singletons of X. An agglomerative hierarchical clustering (AHC) can be summarized in three steps:

1. find A and B such that $d(A, B)$ is minimal.
2. merge A and B in a cluster C.
3. for each remaining cluster D, compute $d(C, D)$.
4. go back to step 1 unless $C = X$.

Differences between algorithms are mainly the way $d(C, D)$ is computed, but steps 1 and 2 can have more than one interpretation. When more than one pair $\{A, B\}$ realize the minimum of d, the choice can be random or lexicographic, or d can be transformed such that the choice has no further consequence [Barthelémy and Guénoche, 1991]. This usually leads to clusters C larger than $A \cup B$. Due to the origin of our data, minimum of d can be considered as unique, and therefore the possible strategies for steps 1 and 2 are equivalent.

Many strategies for computing the distance between the new cluster $C = A \cup B$ and the other clusters have been explored by Lance and Williams [Lance and Williams, 1967], and formalized under the formula:

$$d^p(C, D) = \alpha_A d^p(A, D) + \alpha_B d^p(B, D) + \beta d^p(A, B) + \gamma |d^p(A, D) - d^p(B, D)|$$

Chen [Chen, 1996] restricts the form of α, β and γ in order to explicit the properties of the indexed hierarchy produced by the algorithm. They are functions of three parameters:

$$r_A = \frac{|A|}{|A \cup B|} \qquad r_B = \frac{|B|}{|A \cup B|} \qquad r_D = \frac{|D|}{|A \cup B|}$$

The parameter p is a nonzero real number.

$$d^p(C,D) = \alpha(r_A, r_D)d^p(A,D) + \alpha(r_B, r_C)d^p(B,D) + \beta(r_A, r_B, r_C)d^p(A,B) + \gamma(r_C)|d^p(A,D) - d^p(B,D)|$$

Most usual agglomerative hierarchical classification algorithm can be written under this formalism:

Algorithm	$\alpha(u,w)$	$\beta(u,v,w)$	$\gamma(w)$	p
single linkage	1/2	0	−1/2	1
complete linkage	1/2	0	1/2	1
Ward's method	$\frac{u+w}{1+w}$	$-\frac{w}{1+w}$	0	2

Such an algorithm is called $LW(\alpha, \beta, \gamma, p)$. One should notice that the value of p for single linkage and complete linkage, which is usually 1, can be any nonzero real number. An LW algorithm is said *space-conserving* if

$$\min\{d(A,D), d(B,D)\} \leqslant d(A \cup B, D) \leqslant \max\{d(A,D), d(B,D)\}$$

Single linkage and complete linkage are space conserving. So ultrametrics are fixed points for these algorithms. Ward method is not space-conserving, but it is *space-dilatating*: the dissimilarity produced by the algorithm is greater than the input, and can be different even if the input dissimilarity is an ultramtetric.

To produce an admissible hierarchy indexed by f, the condition $A \subseteq B \implies f(A) \leqslant f(B)$ must be respected. To achieve this goal on any dissimilarity, the LW algorithm must be *monotonic* [Dragut, 2001]:

(i) $\alpha(u,w) + \alpha(1-u,w) + \beta(u, 1-u, w) \geqslant 1$
(ii) $\alpha(u,w) \geqslant 1$
(iii) $\gamma(w) \geqslant \max\{-\alpha(u,w), -\alpha(1-u,w)\}$

Many aggregation functions cannot be written as LW functions, but can be used to produce an indexed hierarchy. It is the case for any *internal* aggregation function. A family of AHC algorithms based on median functions have been studied in [Osswald, 2003].

3 Data

The database contains 26 sonar images provided by the GESMA (Groupe d'Études Sous-Marine de l'Atlantique). Theses images were obtained with a Klein 5400 sonar with a resolution of 20 until 30 cm in azimuth and 3 cm in range. The sea-bottom deep was between 15 m and 40 m.

These 26 sonar images of different sizes (about 92 m width and 92 m to 322 m length) have been segmented in small-images with a size of 64x384 pixels (*i.e.* of approximately 1152 cm × 1152 cm). We have obtained 4003

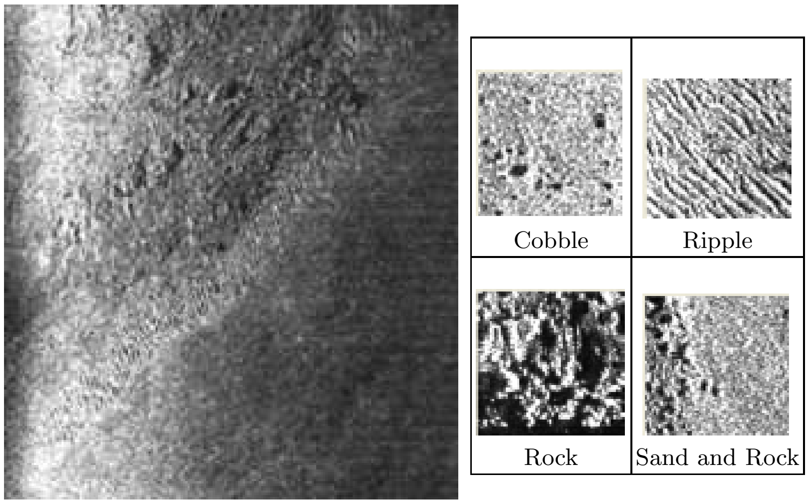

Fig. 1. Sonar image example (provided by the GESMA) and extracted and segmented small-images examples.

small-images. On table 1 we show a sonar image and a sample of these small-images represented in order to obtain a size of 64x64 pixels.

Each small-image is characterized manually by the type of sediment (rock, cobbles, sand, ripple, silt) or shadow when the information is unknown (see Table 1). Moreover the existence of more than one kind of sediment on the small-image is indicated. In this case the type of sediment affected to the small-image is the most present.

Sediment	%	code	% patchworked
Sand	56.06	s	32.00
Rock	19.91	r	43.29
Ripple	9.34	p	61.50
Shadow	8.02	o	47.66
Silt	5.85	i	35.04
Cobble	0.82	c	84.85

Table 1. Percentage and code of type of sediment

From Table 1 we note that the sand sediment is the most represented one. The cobbles sediment is particularly few represented. One of the difficulties of classification step comes from this difference.

There is 38.87% of small-image with more than one kind of sediment (named patch-worked images).

Note that such database is quite difficult to realize. Indeed, the expert has a subjective experience, and can make a mistake for some small-images.

From these small-images, we have extracted texture features. Different texture extraction methods are presented in [Martin *et al.*, 2004]. Each method allow to extract some features that can be redundant, but calculated differently. We choose here to use a wavelet transform.

Indeed, this approach can consider the translation invariance in the directions. The discrete translation invariant wavelet transform is based on the choice of the optimal translation for each decomposition level. Each decomposition level gives four new images on which three features are calculated: the energy, the entropy and a mean. We keep a decomposition level of 3 giving 63 parameters.

So, each small-image is represented in a 63-space. We have calculated the euclidean distance between each small-image: it is the initial dissimilarity used by the AHC algorithms.

4 Usual clustering methods applied on small sonar images

4.1 Some general properties of AHC algorithms

Dissimilarity induced by the single linkage algorithm has the property of being *subdominant*: it is the greatest ultrametric smaller than the original dissimilarity. This constraint often leads to more efficient algorithms [Brucker, 2001]. In the case of ultrametrics, it leads to an algorithm in $\mathcal{O}(n^2)$ operations instead of $\mathcal{O}(n^3)$ for the other LW algorithms.

The single linkage hierarchy is also known to have an *unbalancing effect*: paths from leaves to root have often very different lengths. When A and B are two non-trivial clusters, we also often have $A \subset B$ or $B \subset A$. So it is hard to separate objects into classes: partitions obtained from such a hierarchy are composed of one huge class, and many very small ones.

Other AHC algorithm are not well-defined: applying twice the complete linkage on a dataset may produce two distinct hierarchies, when the dissimilarity d between clusters admits two minimums, and choosing a random one can modify the hierarchy obtained. As our data is composed of floating numbers calculated from real sonar data, the probability of having two minimum in our dissimilarity matrix is nearly 0, so the LW algorithm we use is univocal, and produce binary hierarchies.

4.2 Exemples

Applied to our data, single linkage, complete linkage and Ward algorithm give the trees of figure 2. Index used for the representation is cluster size, for the real index does not allow us to distinguish all the clusters, and the following treatments will only use the clustering structure, not the indices.

We proceed by taking k small-images of each class (k is 4 for examples of figure 2, 12 or 15 for figure 3 data). The proportion of patchworked images, when allowed, is the same than in the original data. As there are only 5 not patchworked cobble images, we consider classes of different size when dealing with larger sets of not patchworked images.

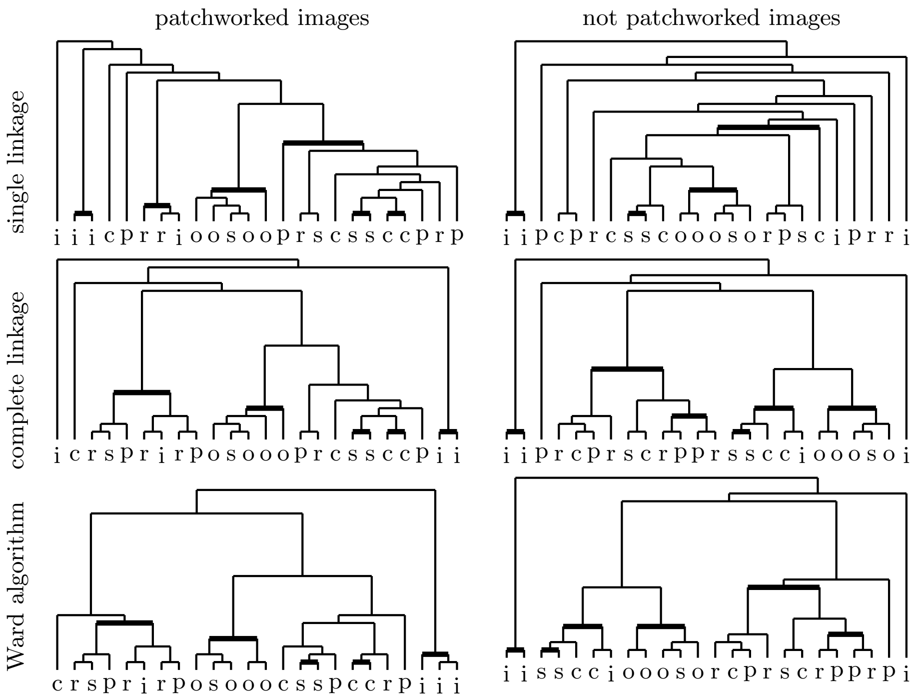

Fig. 2. Usual AHC algorithms applied on some small-images

5 Hierarchy quality

We consider that a hierarchy is efficient for seabed characterization if it contains clusters that are *representative* of each sediment. An expert has defined six classes $M_1, \ldots, M_6$ of small-images, partitionning our data into six sediment classes. We search in the hierarchy $\mathcal{H}$ for clusters A that maximize the quality of an association pattern $A \leftrightarrow M_i$, for i between 1 and 6.

Our concern is how the clusters of the hierarchy can be used as natural clusters for the data. We limit our qualiy measures to the shape of the hierarchy, not its index. A standard (quadratic) distance between the ultrametric induced by the AHC algorithm and the original distance would not help us to reach this goal. As we will see later, Ward's method leads to the most efficient hierarchies, but is space-dilatating. Such a measure would have favored a AHC algorithm between single linkage and complete linkage.

5.1 Measure of hierarchy quality

Tan *et al.* [Tan *et al.*, 2002] have made an exhaustive study of the measures used to measure the quality of association patterns. To obtain a simple measure, depending as little as possible on the size of the dataset, and possible to combine by multiplication, we choose the Jaccard measure, where $P(A)$ is the proportion of elements of A in the dataset:

$$\zeta(A \leftrightarrow M_i) = \frac{P(A \cap M_i)}{P(A) + P(M_i) - P(A \cap M_i)}$$

Combined on a hierarchy, we obtain the quality measure $q(\mathcal{H})$. Bold clusters on figure 2 are the clusters maximizing the ζ measure for at least one type of sediment.

$$q(\mathcal{H}) = \prod_{i=1}^{6} \max_{A \in \mathcal{H}} \zeta(A \leftrightarrow M_i)$$

What is used in the characterization step is not usually a pattern $A \leftrightarrow M_i$ but a rule $A \rightarrow M_i$. As we do not need (and often not want to have) a symmetrical measure for A and M_i, we should use an association rule measure instead of an association pattern measure.

As we want to avoid too small rules, *i.e.* $A \rightarrow M_i$ with $|A| \ll |M_i|$, our measure must take into account the unexplained examples, *i.e.* elements of M_i which are not in A. The Confidence measure ($c((A \leftrightarrow M_i) = 1 - \frac{P(A \cap \overline{M_i})}{P(A)}$) and all the other similar measures are not accurate to achieve this duty (see Vaillant *et al.*, [Vaillant *et al.*, 2004]). The Piatetsky-Shapiro measure, a non-symmetrical extension of the support measure, seem to be the most accurate: $PS(A \rightarrow M_i) = P(A)P(\overline{M_i}) - P(A, \overline{M_i})$ where $\overline{M_i}$ is the complementary of M_i.

5.2 Parameters for Lance-Williams algorithms

Lance and Williams functions associated to single linkage, complete linkage and Ward's method are given section 2.

We build a continuous family of LW algorithms containing those three usual methods. In order to guarantee that $\alpha(u, w) + \alpha(1 - u, w) + \beta(u, 1 - u, w) \geqslant 1$ and therefore that the AHC algorithm obtained is monotonic, we use an intermediary link:

$$\alpha_i(u,w) = \frac{u + w/2}{1+w} \quad \beta_i(u, 1-u, w) = 0 \quad \gamma_i(w) = 0 \quad p_i = 2$$

We use three segments of the space of admissible monotonic *LW* algorithms. The parameter x varies in $[0,1]$.

	$\alpha(u,w)$	$\beta(u,v,w)$	$\gamma(w)$	p
Single to Complete	$1/2$	0	$x - 1/2$	1
Complete to Intermediary	$(1-x)/2+$ $x(u+w/2)/(1+w)$	0	$(x-1)/2$	2
Intermediary to Ward	$(u+(1+x)w/2)/(1+w)$	$-xw/(1+w)$	0	2

We apply this family to random restrictions of our set of small-images, composed of pure small-images or a combination of pure and patchworked small-images. We estimate the efficiency of the *LW* functions on these restrictions.

The quality measure relies on the form of the hierarchy: presence or absence of a cluster. Let $\mathcal{H}(x)$ be the hierarchy produced by the LW algorithm of parameter x. There exists reals x_1 and x_2 such that $x_1 < x < x_2$ and for each $t \in]x_1, x_2[$ we have $\mathcal{H}(t) = \mathcal{H}(x)$. Therefore the quality measure $q(\mathcal{H}(x))$ is locally constant.

On figure 3 we can note the Ward's method is the best *LW* algorithm of the family considered to classify our data. It is not possible to extend the β function joining the intermediary linkage to Ward to x greater than 1, for a value of β lesser than $-w/(1+w)$ would not respect the (i) condition of monotonicity.

5.3 Use of optimized AHC algorithm for texture identification

To use the hierarchy as a characterization tool, we first optimize the *LW* functions on a learning set. We merge this set with small-images whose class is unknown, and we build a hierarchy on this new set, with the same *LW* functions. Then we classify the unknown elements belonging to an optimal class of a sediment type.

We use a set of 72 elements for learning purpose (12 of each sediment type), allowing patchworked small-images, and we add 228 untagged elements. The procedure give us good results for silt and shadow. 100% of small-images tagged by silt are effectivly silt, and 68% for shadow. Among the 228 small-images to classify, 41 received a correct tag, 101 received one

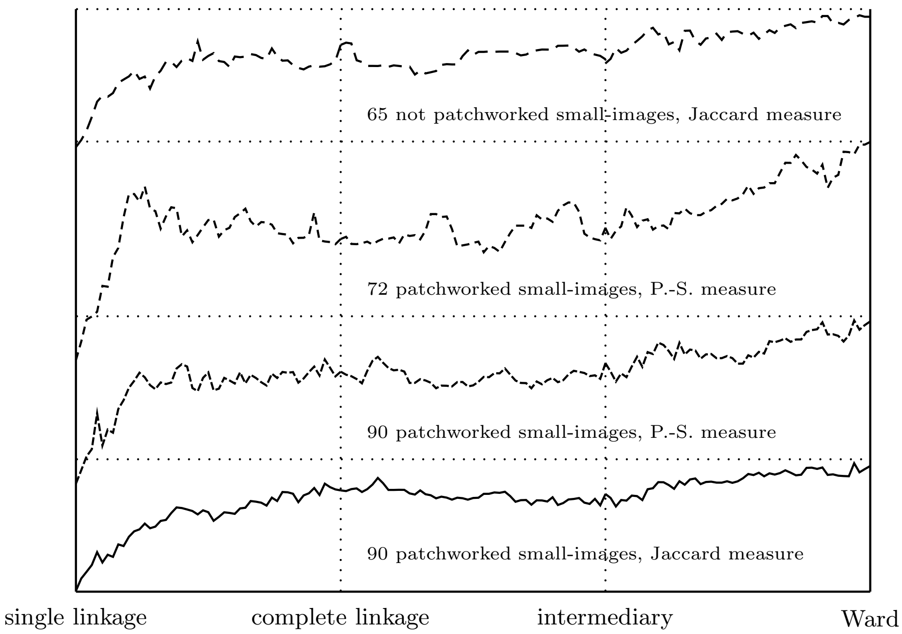

Fig. 3. Hierarchy quality

correct tag and one other tag, 75 received no correct tag and 11 received no tag at all.

Most unclassified small-images are silt (but most silt is well-classified); most ripples small-images are not correctly classified, but Martin *et al.* showed that the wavelets are not an efficient features set to discriminate ripples, as it is not rotation invariant.

6 Conclusion

This approach mixes non-supervised classification methods and supervised classification goals. The supervised context allows us to optimize the AHC parameters, and the tagging method used allow an image to receive one, zero or more than one tag. In a system were several classifiers collaborate, powerful fusion algorithms may use this information.

Here, Ward's method is the most accurate. This may be because of the way dissimilarity is calculated: inertia is closely related to euclidean model. Maybe the fact our classes are of similar size is the origin: Ward's criteria is space-dilating, so it tends to build balanced hierarchies.

References

[Barthelémy and Guénoche, 1991]J.-P. Barthelémy and A. Guénoche. *Trees and Proximity Representations*. Wiley, New York, 1991.

[Brucker, 2001]F. Brucker. *Modèles de classification en classes empiétantes*. PhD thesis, EHESS, 2001.

[Chen, 1996]Z. Chen. Space-conserving agglomerative algorithms. *Journal of classification*, 13:157–168, 1996.

[Dragut, 2001]A. Dragut. Characterization of a set of algorithms verifying the internal similarity. *Mathematical Reports*, 53(3-4):225–232, 2001.

[Jardine and Sibson, 1971]N. Jardine and R. Sibson. *Mathematical Taxonomy*. Wiley, London, 1971. part II.

[Lance and Williams, 1967]G. N. Lance and W. T. Williams. A general theory of classificatory sorting strategies. *The computer journal*, 3-4(9-10):373–380 and 271–277, 1967.

[Martin *et al.*, 2004]A. Martin, G. Sévellec, and I. Leblond. Characteristics vs decision fusion for sea-bottom characterization. In *Colloque Caracterisation in-situ des fonds marins*, Brest, France, 2004.

[Osswald, 2003]C. Osswald. Robustesse aux variations de méthode pour la classification hiérarchique. In *XXXVèmes Journées de Statistiques*, pages 751–754, Lyon, 2003. SFdS.

[Tan *et al.*, 2002]P.-T. Tan, V. Kumar, and J. Srivastava. Selecting the right interestingness measure for association patterns. In *SIGKDD '02*, Edmonton, Canada, 2002.

[Vaillant *et al.*, 2004]B. Vaillant, P. Lenca, and S. Lallich. Association rule interestingness measures: an experimental study. Technical Report 2, GET / ENST Bretagne, 2004.