Biochimie (1993) 75, 379–397
© Société française de biochimie et biologie moléculaire / Elsevier, Paris

Likelihood linkage analysis (LLA) classification method: An example treated by hand

IC Lerman

Irisa-Inria Rennes, CNRS-URA 227, Institut de recherche en informatique et systèmes aléatoires, Campus de Beaulieu, 35042 Rennes Cedex, France

(Received 16 December 1992; accepted 29 December 1992)

Summary — This paper describes a very general method of data analysis using a hierarchical classification. The data can be provided by observation, experiment or knowledge; their nature can be numerical, qualitative or logical. First, the classical view of the context of data representation, in which the algorithm of hierarchical ascendant construction of the classification tree is set, is treated in a synthetic manner. The main notion in our method is one of 'similarity'. This must be elaborated in the best way, taking into account the mathematical nature of the objects to be compared. Here we adopt a set of theoretical and combinatorial representation of the descriptive attributes, which are interpreted in terms of relations. Then we introduce a probability scale for similarity measurement by using a likelihood concept. The largest part of the paper concerns an illustrating example, moderately sized, detailing very minutely the different steps and the different calculations assumed by the method. The data structure handled with this example is the simplest possible. Then, general aspects and methodological extensions are evoked. We end by indicating the interest of the described approach in future works, in which we are involved, concerning typological organization of genetic sequences. We emphasize the 'explanation' aspect of the obtained results, with respect to a given description. For this purpose, classifications (on the object set and on the attribute set) on the one hand and machine learning techniques on the other, intervene efficiently.

hierarchical classification / relational attributes / probabilistic association coefficients / application to genetic sequences

Introduction

In his paper, Saurin [1] emphasized the main interest of multiple comparison of a large family of genetic sequences. Automatic decomposition of this family into classes and subclasses 'significantly' associated, may be crucial in order to induce phylogenies and characterizations. Two aspects can intervene for the latter 'explanation' purpose: functional or structural.

The aim of this paper is presenting a methodology of hierarchical classification based on a probabilistic notion of 'similarity' between combinatorial structures [2, 3]. The extreme generality of this notion and the manner by which it is built enable to take into account *a priori* knowledge in comparison of data units of the set to be organized.

This approach, which refers to combinatorial data analysis [4], employs and develops the algorithmic aspects of hierarchical ascendant construction of a classification tree, by successive agglomerations. However, this approach has not only algorithmic aspects. Its elaboration is at the intersection of three fields: 'combinatorics', 'logic' and 'non-parametric statistics'.

In fact, it gives a very general view of the 'data' and of their automatic synthesis. In this respect, the descriptive 'attributes' (one can also say descriptive 'variables' or 'parameters') are interpreted in terms of relations on the described set. On the other hand, set theoretical and combinatorial representation is adopted for the defined relations (see below).

Additionally, this method introduces an original notion of 'statistics' for measuring statistical relationships and proximities, namely, the 'likelihood' concept. Thus, we set up the 'likelihood' notion as a part of the 'resemblance' notion and we give a probabilistic interpretation of 'similarity', mentioned above. More precisely, association coefficients between descriptive variables (respectively similarity indices between described elementary objects or concepts) will refer to a probability scale for the evaluation of the involved links. This principle also underlies the 'information theory' formalism, in which the higher the amount of information quantity the more unlikely the event concerned. In our case, the events correspond to the observed relations.

A distinctive and important point of the proposed method consists of detection of 'significant' nodes and levels of the classification tree. Intuitively speaking, a 'significant' node corresponds to achievement of a class recovering a concept, at a given degree of

synthesis, while a significant level determines a partition corresponding to an equilibrium state within the clustering synthesis provided by hierarchical classification.

The first chapter is devoted to recall, in a summarized form, the 'classical' view of ascendant hierarchical classification (AHC). Then, the place and the objective of the LLA method will be justified.

Since one goal of this paper is to make explicit the details of the sequence of calculations of the LLA method, the second chapter is necessarily the largest. An example of a small sized data table is considered: 10 objects x 8 attributes, where the attributes are boolean. The concerned mathematical structure is the simplest one for the method. The illustration of the main ideas of this approach will be expressed in the framework of the latter example. Afterwards, general indications will be given to perform the process in the most general data structure, from logical and statistical points of view.

As a matter of fact, the most general data structures – which can be handled by LLA method – are defined in the third chapter. There some extensions or methodological developments related to the approach are also mentioned.

Finally, a brief conclusion will give the general direction of future works that we have begun. Some of them are closely related to conceptual clustering, as defined in machine learning. My strategy will be essentially empirical and deduced from interaction between combinatorial, statistical and computational methods on the one hand and most complex data, provided by the fascinating domain of molecular biology, on the other.

'Classical' view of ascendant hierarchical classification (AHC) and justification of the LLA method

AHC denotes 'hierarchical classification' methods in which the classification is obtained according to an algorithm of ascendant construction by successive agglomerations .

Usually the context of data representation in which such an algorithm is expressed can be set up by means of the following triplet:

$$(O, \mu_O, d) \tag{1}$$

where O is a finite set of elementary objects, provided by a positive measure μ_O assigning to each of its elements x, a weight μ_x. μ_x is associated with the 'importance' with which x has to be considered, $x \in O$. d denotes a dissimilarity or distance index defined on O. Most often, in the classical view, relative to the description of O by descriptive variables, one seeks, in a more or less justified way, to represent the couple (O, μ_O) by a cloud of points in a geometrical space. On the other hand, it is important to provide the representation space with a metric in order to evaluate faithfully the resemblances between objects. Thus, d may be a distance deduced from a metric.

More generally, by assuming the symmetry of the dissimilarity index d (on $O \times O$) and a zero value of d (x, x) for each x belonging to O, we may deduce from (1) the following table:

$$\{d(x,y)), \mu_x, \mu_y/\{x,y\} \in P_2(O)\} \tag{2}$$

where $P_2(O)$ is the set of unordered distinct object pairs.

The characteristic of AHC consists of extending the notion of distance (or dissimilarity) d, between elements of O, to a notion of distance (or dissimilarity) δ between subsets of O. Thus to the triplet (1) we associate the following triplet:

$$(P, \mu_P, \delta) \tag{3}$$

where P is the set of all subsets of O and where μ_P is a positive measure on P, deduced from μ_P.

Let denote the set of real positive numbers. The distance or dissimilarity δ can be put in the following mapping form:

$$\delta : (P \times P, \mu_P) \rightarrow \mathbb{R}_+ \tag{4}$$

Obviously, it is of importance to induce δ from d in a coherent manner, but there does not exist only one construction for this crucial induction. However, formally, we always have:

$$[\forall (X,Y) \in P \times P], \delta(X,Y) = f[\{d(x,y)/(x,y) \in (XUY) \times (XUY)\}, \mu_{XUY}] \tag{5}$$

where the function f is to be defined on the set of mutual distances between weighed elements of $Z = XUY$. On the other hand μ_{XUY} denotes the restriction of μ_o on XUY:

$$(\forall Z \in P), \mu_Z = \{\mu_x/\dot{x} \in Z\} \tag{6}$$

In fact, the definition of δ is only necessary in order to evaluate the distance between two arbitrary disjoint parts of O.

Indeed, the algorithm of ascendant hierarchical classification, using δ to build a classification tree on O, is a 'trivial' mathematical principle: 'at each step, join the class pairs which realise the minimum value of δ'. However, this mathematical 'triviality' does not entail computational and statistical 'trivialities'.

Example

Let $O = \{1,2,3,4,5,6\}$ be the object set, in which each interger number is a symbol which denotes a single object. A classification tree A(O) on O may have the form shown in figure 1.

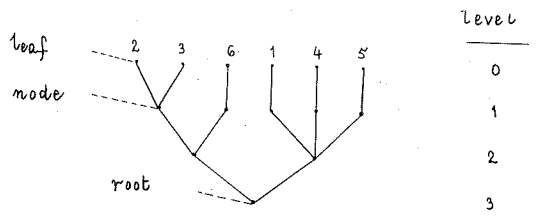


Fig 1. Classification tree.

For the algorithm implementation, we may compare the hierarchical ascendant construction of a classification tree to evolution of a system. If k is a tree level, we characterize the state of the system by the couple (T_k, μ^k) where T_k is the table of the δ indices between the classes formed at the level k and where μ^k is the measure on the set of these classes (compare with equation (3) above). The initial state (T_0, μ^0) is defined by the matrix T_0 of the δ indices between singleton classes, comprising each exactly one element, and by μ^0, where the weight of a singleton class is that of the element concerned. By denoting l the number of the last level of the classification tree, corresponding to the root ($l = 3$ in fig 1), the states of the system for $0 \leq k \leq l-1$, have to be considered. Under these conditions, it is of first importance from computational point of view to have a formula, called the 'reactualization formula', of the following form:

$$(T_{k+1}, \mu^{k+1}) = \varphi(T_k, \mu^k) \qquad (7)$$
$$(a) \qquad (b)$$

(a) state of the system at the level $k + 1$

(b) state of the system at the level k

$0 \leq k \leq l - 2$, where φ is a function to be determined with respect to the dissimilarity δ (see eqn (3)).

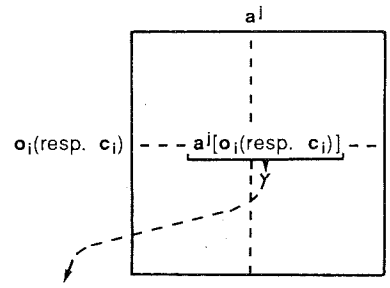
Designating by c_k the number of classes emerged at the level k of the classification tree, the necessary dimension of T_k is c_k $(c_k - 1)$ /2. Indeed, δ is symmetrical and it is not relevant to define the value of δ between one class and itself.

Equation (7) gives the system state at the level $(k + 1)$ after joining the nearest classes of the k-th level on the basis of T_k.

AHC can be a very general and powerful tool for data analysis of a data table $\mathcal{T}$ corresponding to a description. For this purpose, we have to specify the general structure of $\mathcal{T}$. In fact, there are two fundamental cases for the latter structure, which can be denoted: 1) O x A; and 2) C x A, where the set of rows of $\mathcal{T}$ is labelled by O (respectively C) and where the set of columns of $\mathcal{T}$ is labelled by A.

O represents a set of elementary objects (one can also say 'individuals'). We mean that each element of O cannot be divided, while C defines a set of classes (one can also say 'concepts'). In both cases (1) and (2), A is a set of descriptive attributes (or 'variables').

Let us consider the following diagram which represents the data table $\mathcal{T}$:



Observation or knowledge interpreted in a logical-statistical way.

Fig. 2. General structure of a data table.

Denote by n and p respectively, the number of rows and the number of columns of the data table $\{n =$ card (O) [respectively card (C)] and $p =$ card $(A)\}$. Thus O (respectively C) and A can be denoted as follows:

$$O = \{o_i / 1 \leq i \leq n\} \text{ (respectively } C = c_i / 1 \leq i \leq n\})$$

and

$$A = \{a^j / 1 \leq j \leq p\}$$

The entry which is at the intersection of the i^{th} row, labelled by the object o_i (respectively the concept c_i) and of the j^{th} column, labelled by the attribute a^j, contains the 'value' of a^j on o_i (respectively c_i), $1 \leq i \leq n$, $1 \leq j \leq p$. But this 'value' notion has not only to be understood relatively to the classical definition of the couple (attribute, value) as is the case in [6]. a^j $[o_i$ (respectively $c_i)]$, which is obtained either from observation or knowledge, may correspond to a modal logical formula on the set of 'categories' (one can also say 'modalities' or 'values') underlying the measure scale of the variable a^j. In our approach statistical-logical interpretation of a^j $[o_i$ (respectively $c_i)]$ is considered, $1 \leq i \leq n$, $1 \leq j \leq p$.

We have already expressed above the classical context of the general presentation of AHC. The latter approach forgets the following two fundamental points: i) A comprises qualitative variables, where the set of the 'values' ('modalities' or 'categories') of a given variable is (very) structured by the domain or expert knowledge; and ii) the specific interest of the hierarchical classification of the variable set A into 'significant' classes and subclasses. Such a decompo-

sition provides a very interesting alternative to factorial analysis. In fact, it enables the discovery of a system of factors and subfactors; the subfactors of a factor are 'relatively independent' in the context of the factor.

Consider i). Even description of the data by quantitative variables has directly to be considered, the significant descriptive information, contained in a given variable, for classificatorial goal, is better set up if we transform this variable to a qualitative one. This transformation is obtained by means of an optimal subdivision, taking into account expert knowledge, of the total range variation and by associating with each interval a modality (category). The whole set of these modalities is the value set of the qualitative variable obtained. The value set can be structured in an ordinal way (ordinal qualitative variable or pre-ordonance variable) [7–9].

Considering point ii), it is clear that the set E to be classified can either be the set O of objects (respectively C of concepts) or the set A of descriptive attributes:

E: set to be classified $\quad E = A$

$\qquad\qquad\qquad\qquad E = O$ (respectively C)

Fig 3. Alternatives in classification.

The reason why this double classification can be processed, by respecting i) above, is precisely the simplicity of the AHC algorithm principle. Indeed, there are considerable scientific advantages, *ie* the flexibility of use enables us to take into account multiple types of data description, but also dangers, *ie* the 'significance' of automatic synthesis depends on two crucial aspects: i) the 'relevant coding' of information, results from experimental data or knowledge data; and ii) the 'relevant notion of proximity' on the set $P(E)$ of all subsets of the set E to be classified.

Aberrant results can be obtained by AHC if the previous aspects i) and ii) are not considered very minutely.

As we have implicitely mentioned above the 'likelihood linkage analysis (LLA)' hierarchical classification method addresses any logical-mathematical type of the data table $\mathcal{T}$ (see fig 2). The general proceeding of the LLA method can be stated as follows:

$$DO: E = A$$
$$ARBA = LLA(E)$$
$$DO: E = \text{(respectively } C)$$

according to the nature of the data table

$$ARBO = LLA(E).$$

If $E = A \Rightarrow$ set dual $(E) = O$ (respectively C) according to the nature of the data table. If

$$E = O \text{ (respectively } C) \Rightarrow \text{set dual } (E) = A.$$

A very important stage consists of the class 'explanation'. The latter assumes that we are able to 'situate', by means of association coefficients, an organized system of classes on a subset of E (respectively dual (E)), respecting an organized system of classes on a subset of dual (E) (respectively E).

Under these conditions, one can for example speak of a notion of the 'responsibility degree' of such a class of attributes, in the forming of such an object class.

The management of the relative position of *ARBA* with respect to *ARBO* (*cf* above) enables artificial intelligence techniques to intervene with much more efficiency in the explanation of automatic synthesis provided by the classification tree.

The LLA classification method in the simplest case: An example treated by hand

The 'simplest' case for a given method involves a data structure in which the elaboration of the method is the easiest to be expressed, here the case of a set A of boolean attributes (also called presence-absence variables). On the other hand, the description concerns a set O of elementary objects (see section above).

Under these conditions and as announced in the *Introduction*, we are going to illustrate the previous processing diagram in the following, very simple example of an incidence data table comprising 10 rows and eight columns (see table I below). The objects (respectively the attributes) are represented by rows (respectively by columns).

Table I. Data table

O	A a^1	a^2	a^3	a^4	a^5	a^6	a^7	a^8
o_1	0	0	1	1	0	1	0	1
o_2	0	0	0	1	0	0	1	1
o_3	1	1	0	1	1	0	0	0
o_4	0	0	1	0	0	0	1	1
o_5	1	0	1	0	1	0	0	0
o_6	1	0	0	1	1	0	0	0
o_7	1	0	0	0	1	1	1	0
o_8	0	0	1	1	0	0	1	1
o_9	1	0	0	1	1	1	0	0
o_{10}	1	1	0	1	0	0	1	1

Association between two attributes

A given presence-absence descriptive attribute a defines an unary relation on the object set O (for more details

see the following section *Methodological extensions*). a is represented by the subset $O(a)$ of O, consisting of all objects where the attribute is present (one also says: where the boolean attribute is 'true').

Relative to a pair $\{a, b\}$ of boolean attributes, the following parameters which represent set cardinalities are introduced. $\tilde{a}$ and $\tilde{b}$ denote respectively the complemented variables of a and b.

$$\left.\begin{aligned} s &= \text{card } [O(a) \cap O(b)] \\ u &= \text{card } [O(a) \cap O(\tilde{b})] \\ v &= \text{card } [O(\tilde{a}) \cap O(b)] \\ \text{and} \quad t &= \text{card } [O(\tilde{a}) \cap O(\tilde{b})] \end{aligned}\right\} \tag{8}$$

Consider boolean attributes resulting from conjunction between two attributes, where the former belongs to $\{a, \tilde{a}\}$ and where the latter belongs to $\{b, \tilde{b}\}$: $a \wedge b$, $a \wedge \tilde{b}$, $\tilde{a} \wedge b$ and $\tilde{a} \wedge \tilde{b}$. By associating with these attributes their respective extensions at the level of the object set O, the preceding cardinalities (see eqn (8)) can take place in the following diagram:

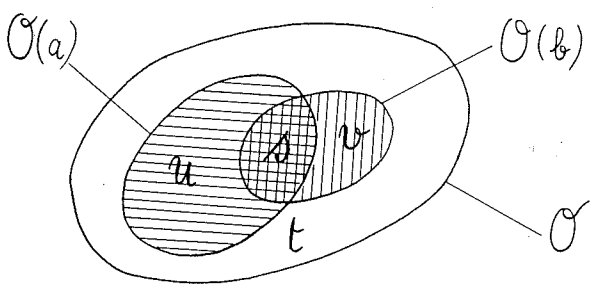


Fig 4. Comparison of two subsets of the object set.

It is obvious that the statistics s (number of objects which possess the two attributes a and b) must play an important part in the construction of the similarity measure. As a matter of fact, the presence of two features in the same object can be significant for their association.

$$s(a, b) = \text{card}[O(a) \cap O(b)] \tag{9}$$

will be called 'raw' index of similarity. But the value of s alone is certainly a biased index of the resemblance between the two attributes a and b; to obtain a rather high (or low) value of s it is indeed sufficient to have two frequently-occurring (or rare) features, irrespective of the relative position of $O(a)$ and $O(b)$. Therefore, it is in question to neutralize the size influence of $O(a)$ and $O(b)$, *ie* card $[O(a)]$ and card $[O(b)]$. It is exactly what is tried, in an implicit manner, by many similarity indices conceived to compare rows or columns of an incidence data table composed of zeros and ones. In fact each of these indices can be expressed in terms of a particular function $I(s, u, v)$ of the triplet (s, u, v), increasing with respect to s, symmetrical with respect to u and v and decreasing with respect to u; on the other hand, the increasing with respect to s or the decreasing with respect to u, is strict ($s \geq 0$, $u \geq 0$, $v \geq 0$ and $s + u + v \leq n$).

More precisely, the general form of each of the proposed functions $I(s, u, v)$, is a ratio. The numerator and the denominator of the latter are generally rather simple functions, utilising elementary operations (addition, subtraction, multiplication, division and square root), on the following parameters: s, $s + u = n(a) = \text{card } [O(a)]$ and $s + v = n(b) = \text{card } [O(b)]$. Thus, the famous Jaccard index [10] can be expressed as follows:

$$J(a,b) = \frac{s}{s+u+v} = \frac{s}{(s+u)+(s+v)-s} = \frac{\text{card}[O(a) \cap O(b)]}{\text{card}[O(a) \cup O(b)]} \tag{10}$$

a and b are assumed to belong to the attribute set A. But consider here the level of comparing a and b, independently of the context of the other attributes. The proposed approach to elaborate a probabilistic index between a and b consists of associating with the couple $[O(a),O(b)]$ of observed subsets of O, a couple (X, Y) of independent random subsets of a set Ω to be defined. In this random association, the triplet of cardinalities $[n(a),n(b),n]$ is respected in a probabilistic manner by the triplet [card (X), card (Y), card (Ω)] of integer random variables, respectively associated (we will be more accurate on this point below). The correspondence

$$[O(a), O(b), O] \longrightarrow [X, Y, \Omega] \tag{11}$$

is what we call 'hypothesis of no link (h.n.l.)' or more commonly 'hypothesis of independence'.

In this framework consider the correspondence

$$s = \text{card } [O(a) \cap O(b)] \longrightarrow S = \text{card } (X \cap Y) \tag{12}$$

associating with the observed raw index s, the random raw index S. Under these conditions, the greater the resemblance measure between the two attributes a and b, the more unlikely the size of the s value is. More precisely, resemblance measure is a decreasing function of the probability

$$Pr\ \{S > s/h.n.l.\} \tag{13}$$

calculated from the probability distribution of S.

Then, to evaluate the similarity between the two attributes a and b, one can take the complementary probability of (13), namely:

$$P(a, b) = Pr\ \{S \leq s/h.n.l.\} \tag{14}$$

By taking $P(a, b)$ as association coefficient between a and b, we set up the 'likelihood' concept as a part of the 'resemblance' concept. Once more that resemblance measure is established by reference to random hypothesis of 'uniform perfect chance', as figured in (11). Nevertheless, there are three fundamuntal versions for the h.n.l. [2]. The variation concerns on the one hand the choice of Ω and on the other the random model in order to select X (respectively Y) in the set $P(\Omega)$ of all subsets of Ω. Whatever the considered random model, by designating by E mathematical expectation, we have:

$$\left.\begin{aligned} &E[\mathrm{card}(\Omega)] = n \\ &E[\mathrm{card}(X)] = n(a) = \mathrm{card}[O(a)] \\ &E[\mathrm{card}(Y)] = n(b) = \mathrm{card}[O(b)] \end{aligned}\right] \tag{15}$$

These relations mean that the random model respects the cardinalities of the compared structures. In fact, for each of the two first models (among the three models), we have $\Omega = O$. Some aspects of the second model will be considered for the following calculations.

For this model, X (respectively Y) is a random element in the set $P(O)$ of all subsets of O. For the random choice of X (respectively Y), $P(O)$ is provided by a probability measure P_a (respectively P_b), such that the probability to obtain for X (respectively Y) a subset of which the cardinality is l (respectively m), is given by:

$$p(a)^l p(\tilde{a})^{n-l}[\mathrm{resp}\ p(b)^m p(\tilde{b})^{n-m}] \tag{16}$$

where respectively $p(a)$, $p(\tilde{a})$, $p(b)$ and $p(\tilde{b})$, the proportions $n(a)/n$, $n(\tilde{a})/n$, $n(b)/n$ and $n(\tilde{b})/n$ are denoted. Under these conditions, it is proven [2] that the random variable S (see (12)) is binomial with the parameters n and $p = p(a) \times p(b)$. Therefore, the above index (14) can be written:

$$P(a, b) = \sum_{0 \leq i \leq s} \binom{n}{i} p^i (1-p)^{n-i} \tag{17}$$

We will have to apply this index for the considered example (see table I above), where $n = 10$. Obviously, since n is small, direct calculation of (17) can be handled. It was preferred to present below the table of the cumulative function of binomial distribution $B(n = 10, p)$, for different values of the parameter p.

By consulting this table, one obtains for example:

$$P(a^3, a^7) = 0.678 \text{ and } P(a^4, a^7) = 0.514$$

More completely, consider table IV where the values of the probabilistic indices are established from table III and by using table II:

$$\{P(a^j,a^k)/1 \leq j < k \leq 8\} \tag{18}$$

Beside the table of probabilistic indices, consider the table of 'normalized' (or 'standardized') indices, where the raw index is 'centered and reduced' with respect to 'hypothesis of no link'. Denoting by:

$$\{Q\ (a^j,a^k)\ /1 \leq j < k \leq 8\} \tag{19}$$

the latter table, we have:

$$Q(a^j, a^k) = \frac{s(a^j, a^k) - E[S(a^j, a^k)]}{\sigma[S(a^j, a^k)]} \tag{20}$$

where E and σ^2 designate, respectively, mathematical expectation and variance, $1 \leq j < k \leq 8$ (see table V).

The construction of this mathematical table becomes necessary to evaluate probabilistic indices, if n is too large. In this case, we generally have the excellent approximation given by:

$$P\ (a^j,a^k) \approx \phi\ [Q\ (a^j,a^k)] \tag{21}$$

$1 \leq j < k \leq 8$, where ϕ is the normal cumulative distribution function.

By considering the first random model of independence, the coefficient $Q(a, b)$ is, to multiplicative factor $\sqrt{n-1}$, nothing else than the well known K Pearson coefficient. On the other hand, the third random model leads for $Q(a, b)$, to multiplicative factor $\sqrt{n}$, to what is called: the 'oriented' contribution of (a, b) to the index χ^2/n, associated with the 2 x 2 contingency table, crossing $\{a, \tilde{a}\}$ with $\{b, \tilde{b}\}$.

Now two associations are compared: between a^1 and a^6 on the one hand and between a^4 and a^7 on the other hand. In figure 5 we have represented the whole set O by an horizontal rectangle. A given attribute a^j is depicted by a subrectangle bound by two vertical segments. The length of the subrectangle representing an attribute a^j is proportional to the number of points where a^j takes the value 1; that is:

$$\mathrm{card}[(a^j)^{-1}(1)] = \mathrm{card}[O(a^j)] = n(a^j)$$

We have vertically (respectivelly horizontally) hatched the rectangles representing a^1 and a^4 (respectivelly a^6 and a^7). On the other hand, the rectangle size resulting from the intersection of the two rectangles which represent a^1 and a^6 (respectively a^4 and a^7), respects the cardinality of $O\ (a^1) \cap O\ (a^6)$ [respect-

Table II. Binomial cumulative distribution function $B(n = 10, p)$; $F(k) = Pr\ (K \leq k)$.

n	k	$P = 0.05$	0.10	0.15	0.20	0.25	0.30	0.35	0.40	0.45
10	0	0.5987	0.3487	0.1969	0.1074	0.0563	0.0282	0.0135	0.0060	0.0025
	1	0.9139	0.7361	0.5443	0.3758	0.2440	0.1493	0.0860	0.0464	0.0233
	2	0.9885	0.9298	0.8202	0.6778	0.5256	0.3828	0.2616	0.1673	0.0996
	3	0.9990	0.9872	0.9500	0.8791	0.7759	0.6496	0.5138	0.3823	0.2660
	4	0.9999	0.9984	0.9901	0.9672	0.9219	0.8497	0.7515	0.6331	0.5044
	5	1.000	0.9999	0.9986	0.9936	0.9803	0.9527	0.9051	0.8338	0.7384
	6	1.000	1.000	0.9999	0.9991	0.9965	0.9894	0.9740	0.9452	0.8980
	7	1.000	1.000	1.000	0.9999	0.9996	0.9984	0.9952	0.9877	0.9726
	8	1.000	1.000	1.000	1.000	1.000	0.9999	0.9995	0.9983	0.9955
	9	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.9999	0.9997
	10	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000

Table III. Table of raw indices between descriptive attributes (including and below the main diagonal). Table of the values of the parameter p of the binomial laws to be consulted (strictly above the main diagonal).

	a^1	a^2	a^3	a^4	a^5	a^6	a^7	a^8
a^1	6	0.12	0.24	0.42	0.30	0.18	0.30	0.30
a^2	2	2	0.08	0.14	0.10	0.06	0.10	0.10
a^3	1	0	4	0.28	0.20	0.12	0.20	0.20
a^4	4	2	2	7	0.35	0.21	0.35	0.35
a^5	5	1	1	3	5	0.15	0.25	0.25
a^6	2	0	1	2	2	3	0.15	0.15
a^7	2	1	2	3	1	1	5	0.25
a^8	1	1	3	4	0	1	4	5

Table IV. (first aggregation): above the diagonal: table of the probabilistic indices of 'likelihood of the link', based on the binomial model. Below the diagonal: table of the function 4 $\{-\log_e(-\log_e(\cdot))\}$ values of the probabilistic indices.

	a^1	a^2	a^3	a^4	a^5	a^6	a^7	a^8
a^1	–	0.886	0.270	0.582	0.953	0.735	0.389	0.149
a^2	2.112	–	0.449	0.842	0.736	0.549	0.736	0.736
a^3	–0.270	0.222	–	0.440	0.376	0.659	0.678	0.879
a^4	0.614	1.760	0.197	–	0.514	0.647	0.514	0.752
a^5	3.034	1.182	0.022	0.407	–	0.820	0.244	0.056
a^6	1.178	0.511	0.875	0.831	1.617	–	0.544	0.544
a^7	0.041	1.182	0.945	0.407	–0.344	0.496	–	0.922
a^8	–0.644	1.182	2.048	1.255	–1.059	0.496	2.511	–

Table V. Standardized indices with respect to the binomial model of the h.n.l.: $\{Q\ (a^j,a^k)\ /1 \leq j < k \leq 8\}$.

	a^1	a^2	a^3	a^4	a^5	a^6	a^7	a^8
a^1								
a^2	0.778							
a^3	–1.037	–0.933						
a^4	–0.128	0.547	–0.563					
a^5	1.380	0.000	–0.791	–0.331				
a^6	0.165	–0.799	–0.195	–0.078	0.443			
a^7	–0.690	0.000	0.000	–0.331	–1.095	–0.443		
a^8	–1.1380	0.000	0.791	0.331	–1.826	–0.443	1.095	

ively $O\ (a^4) \cap O\ (a^7)$]. It is not easy to evaluate intuitively, on the basis of visual perception, among $\{a^1, a^6\}$ and $\{a^4, a^7\}$, which is the pair comprising most similar components. The Jaccard index gives:

$$J(a^1, a^6) = \frac{2}{7} = 0.286 < J(a^4, a^7) = \frac{3}{9} = 0.333$$

This inequality is indeed reversed for the 'likelihood of the link' index. Explicitely, we have:

$$P\ (a^1, a^6) = 0\ 737 > P\ (a^4, a^7) = 0.514$$

Thus, for the P index, an intersection size greater than 2 and concerning two subsets of which the cardinalities are respectively 3 and 6, is more unlikely than an intersection size greater than 3 and concerning two subsets of which the cardinalities are respectively 5 and 7.

Let us indicate that the numbers of the table IV, situated above the main diagonal, are exclusively obtained on the basis of table II. Linear interpolation is considered when the value of the parameter p is not included.

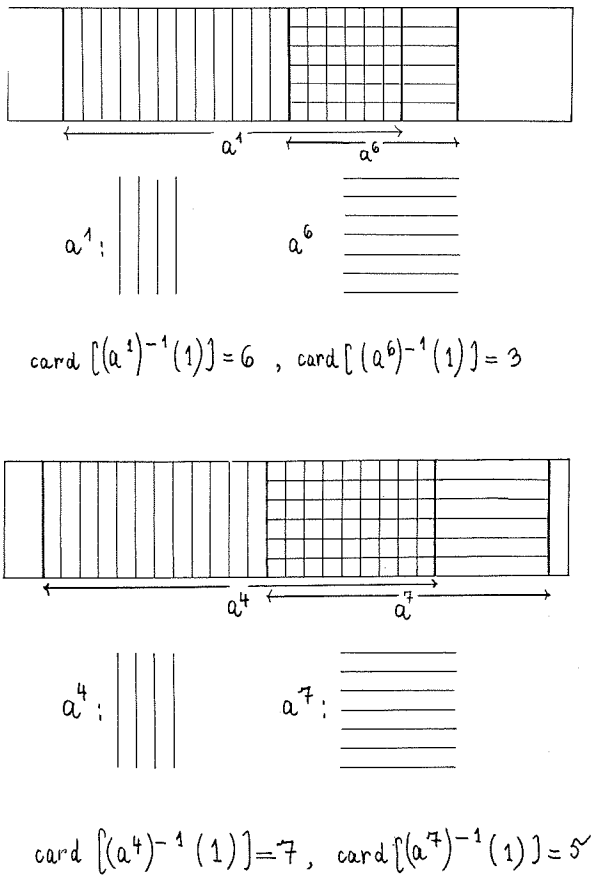


Fig 5. Comparing two associations.

From this data structure concerned with incidence table, generalizations may either deal with statistical aspects (the number n of rows of the data table is more or less large), or with formal aspects related to the structure underlying the category set of relational variables.

Before introducing the family of criteria of the 'maximal link likelihood' (see next section), it has made to be clear that whatever is the set E on which classificatory organization has to be discovered [$E = A$ or $E = O$ (respectively C) (see fig 2)], it always results in probabilistic similarity indices table, as suggested in the following diagram:

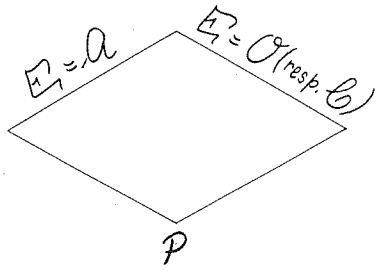


Fig 6. Common probalistic structure of the similarity indices.

$$\mathcal{P} = \{P(x, y) / \{x, y\} \in P_2(E)\} \qquad (22)$$

Now consider the probabilistic independence hypothesis (or h.n.l.) between the respective attributes. It is defined by the correspondence

$$A \rightarrow A^* \qquad (23)$$

where A^* is a family of independent random attributes, respectively associated to the attributes belonging to A, by respecting statistical and formal structures of the latter attributes, taken one by one.

The function (22) is established in such a way that if $\{x^*,y^*\}$ is the pair of the two independent random elements associated with $\{x,y\}$ of $P_2(E)$, according to the above random model (see (23)), then $P(x^*,y^*)$ is a (0, 1) uniform random variable, whatever is the pair $\{x,y\}$ of $P_2(E)$.

Family of criteria of the 'maximal link likelihood'

In order to understand in an intuitive manner this type of proximity criterion between classes, consider figure 7, which is placed in the simple and classical context of a two-dimensional cloud. We suggest two classes of pairs: $\{C_1, C_2\}$ on the one hand and $\{D_1, D_2\}$ on the other. In this drawing, the shaded density expresses the importance of the density. Thus, C_1 and C_2 are strongly dense, but D_1 and D_2 are more weakly dense.

Let us begin by considering a 'distance' index δ between the subsets of a metrical space, provided by topology:

$$\delta(x, y) = \min \{d(x, y) / (x, y) \in X \times Y\} \qquad (24)$$

where X and Y are two parts of the space concerned.

We have:

$$\delta(C_1, C_2) < \delta(D_1, D_2) \qquad (25)$$

In these conditions, do we have to join C_1 and C_2 before D_1 and D_2 ? In fact we consider δ as a 'raw'

index. The likelihood of the maximal link criterion does the opposite of the previous proposition: it begins by aggregating D_1 and D_2 before aggregating C_1 and C_2; because the smallness of $\delta\ (D_1,D_2)$ is more exceptional than that of $\delta\ (C_1, C_2)$ by considering the point densities of the two classes to be compared.

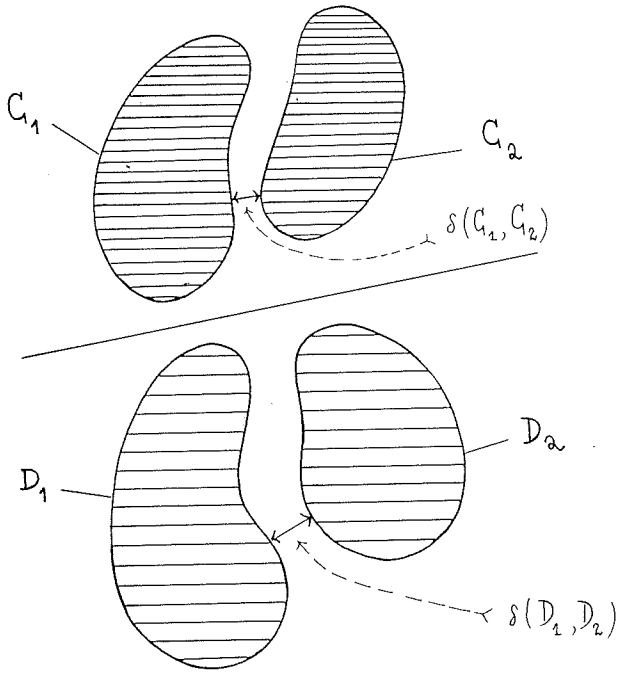


Fig 7. Comparing two class dissimilarities.

This idea also appears fundamental in information theory, where the higher the quantization of an event, the more unlikely it is. The events with which we are concerned here are the observed relations between descriptive variables (respectively described objects or concepts), or between variable classes (respectively object classes or concept classes).

Reconsidering the last passage, just before this section, the starting point of the construction of the aggregation criterion between two classes C and D (disjoint parts of E) which is called the 'maximal link likelihood' is:

$$p(C, D) = \max\ \{P(c, d) / (c, d) \in C \times D\} \quad (26)$$

To the couple (C, D), we associate a couple (C^*, D^*) of independent random classes, where C^* (respectively D^*) is composed of independent elements related to the statistical structure of the C-elements (respectively D-elements). This association corresponds to the 'hypothesis of no link (h.n.l.)' (or 'independence') H described above (see previous section and (23)] where to the observed attributes we associate independent attributes, respecting their combinatoric types.

Thus $p(C^*,D^*)$ becomes a 'random index'. Then the definitive association index between the two classes C and D takes the following form:

$$P(C, D) = Pr\ \{p\ (C^*, D^*) \leq p\ (C, D) / H\} = [p\ (C, D]^{l \times m} \quad (27)$$

where l = card (C) and m = card (D).

The latter index corresponds to what we call the pure form of the maximal link likelihood criterion, that we can denote by:

$$V L_1\ (C, D) = [p\ (C, D)]^{l \times m} \quad (28)$$

whereas $p(C, D)$ (see (26) above) corresponds to the maximal link [2].

Following the work of Nicolaü [11], we have suggested the family of criteria:

$$V L_\varepsilon\ (C, D) = [p\ (C, D)]^{(l \times m)^\varepsilon} \quad (29)$$

where ε is a real parameter between 0 and 1. This family goes from the maximal link ($\varepsilon = 0$) to the pure form of the maximal link likelihood criterion ($\varepsilon = 1$).

For reasons of accuracy in computing, we consider the strictly increasing function:

$$S_\varepsilon(C, D) = -\log\{-\log[V L_\varepsilon(C,D)]\} \quad (30)$$

which leads to exactly the same classification tree.

We do have a reactualization for S_ε in case of multiple aggregation:

$$S_\varepsilon(C, D) = -\varepsilon \log(\sum_{1 \leq i \leq k} c_i) - \varepsilon \log(\sum_{1 \leq j \leq l} d_j) + \max\{S_\varepsilon(C_i, D_j) + \varepsilon \log(c_i) + \varepsilon \log(d_j) / 1 \leq i \leq k, 1 \leq j \leq l\} \quad (31)$$

where we have denoted by c_i = card (C_i), $1 \leq i \leq k$ (respectively d_j = card (D_j), $1 \leq j \leq l$) and where $C = C_1 \cup C_2 \cup \ldots \cup C_k$ (respectively $D = D_1 \cup D_2 \cup \ldots \cup D_l$).

However, the formula that we will employ below has the following classical form:

$$S_1(C, D_1 \cup D_2) = -\log(d_1 + d_2) + \max\{S_1(C, D_1) + \log(d_1), S_1(C, D_2) + \log(d_2)\} \quad (32)$$

which is associated with the value 1 of ε. This formula reactualizes the similarities between classes after aggregation of only two classes.

The tables IV, VI–XI give the sequence of the states of the couple designated by (T_k,μ^k) in the first chapter, $0 \leq k \leq 6$. Table IV is associated with (T_0,μ^0) and the table XI with (T_6,μ^6). Therefore, by including the zero level (see fig 8), the obtained tree comprises eight levels. The weighting μ^k is defined here by the

class cardinalities. These integers take place in the last column at the right side (see tables VI–XI).

For each similarity matrix T_k, $0 \leq k \leq 6$, we have underlined the maximum value, indicating by this way the couple of classes to merge. In our example, for each step, there is only one couple of classes which realizes the latter maximum value of the similarity index between classes. Thus the tree is binary (see fig 8). This particular case is directly related to the example, but does not correspond to a general fact.

When k increases, $0 \leq k \leq 6$, the number of formed classes decreases. We have, in the following tables (VI–XI), clearly indicated the evolution of a given labelled class, according to the column (respectively row) concerned in the matrix T_k, $0 \leq k \leq 6$. This way, it is possible to retreive the different elements included in a given class which intervenes in a given table T_k, $0 \leq k \leq 6$.

Table VI. (2nd aggregation). Table of the class proximities. The class cardinalities are represented in the last column.

1 ← 1 V 5	*2 ← 2*	*3 ← 3*	*4 ← 4*	*5 ← 6*	*6 ← 7*	*7 ← 8*	
							2
1.419							1
–0.617	0.222						1
-0.079	1.760	0.197					1
0.924	0.511	0.875	0.831				1
–0.652	1.182	0.945	0.407	0.496			1
–1.337	1.182	2.048	1.255	0.496	2.511		1

Table VII. (3rd aggregation). Table of the class proximities. The class cardinalities are represented in the last column.

1 ← 1	*2 ← 2*	*3 ← 3*	*4 ← 4*	*5 ← 5*	*6 ← 6 V 7*	
						2
1.419						1
–0.671	0.222					1
–0.079	1.760	0.197				1
0.924	0.511	0.875	0.831			1
–0.1345	0.489	1.355	0.562	–0.197		2

Table VIII. (4th aggregation). Table of the class proximities. The class cardinalities are represented in the last column.

1 ← 1	*2 ← 2 V 4*	*3 ← 3*	*4 ← 5*	*5 ← 6*	
					2
0.726					2
–0.671	–0.471				1
0.924	0.138	0.875			1
–1.345	0.131	1.355	–0.197		2

Table IX. (5th aggregation). Table of the class proximities. The class cardinalities are represented in the last column.

1 ← 1	*2 ← 2*	*3 ← 3 V 5*	*4 ← 4*	
				2
0.726				2
–1.751	–0.537			3
0.924	0.138	–0.224		1

Table X. (6th aggregation). Table of the class proximities. The class cardinalities are represented in the last column.

1 ← 1 V 4	*2 ← 2*	*3 ← 3*	
			3
0.32			2
–1.323	–0.537		3

Table XI. (7th aggregation). Table of the class proximities. The class cardinalities are represented in the last column.

1 ← 1 V 2	*2 ← 3*	
		5
–1.453		3

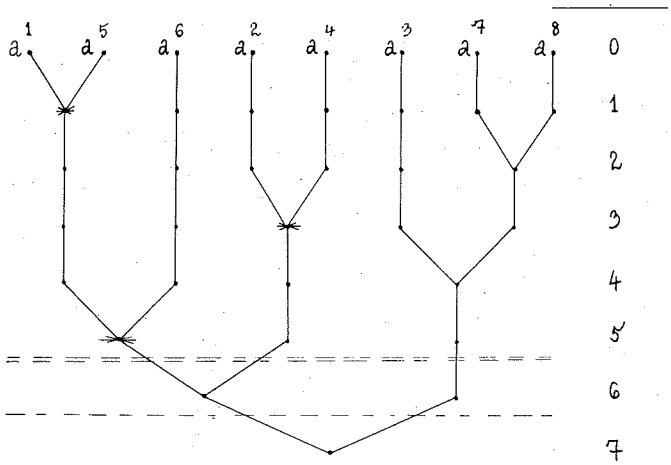


Fig 8. Classification tree on the attribute set.

Associate with the tree levels the sequence of the respective partitions, ordered by decreasing fineness:

$$\begin{aligned}
\pi_0 &= \{\{1\},\{2\},\{3\},\{4\},\{5\},\{6\},\{7\}\{8\}\} \\
\pi_1 &= \{\{1,5\},\{2\},\{3\},\{4\},\{6\}\{7\}\{8\}\} \\
\pi_2 &= \{\{1,5\}\{2\},\{3\}\{4\}\{6\},\{7,8\}\} \\
\pi_3 &= \{\{1,5\},\{2,4\},\{3\},\{6\},\{7,8\}\} \\
\pi_4 &= \{\{1,5\}\{2,4\},\{3,7,8\},\{6\}\} \\
\pi_5 &= \{\{1,5,6\},\{2,4\},\{3,7,8\}\} \\
\pi_6 &= \{\{1,2,4,5,6\},\{3,7,8\}\} \\
\pi_7 &= \{\{1,2,3,4,5,6,7,8\}\}
\end{aligned} \qquad (33)$$

Consider the set $F = P_2(A)$ of unordered element pairs of A. A partition π of A determines a partition of F into two complementary classes denoted by $R(\pi)$ and $S(\pi)$. $R(\pi)$ is the set of joined pairs by the partition π; respectively, $S(\pi)$ is the set of separated pairs by the partition π.

For the above partitions (see (33)), by designating xy the pair $\{a^x, a^y\}$, where $x < y$, we have:

$$
\begin{aligned}
&R(\pi_0) = \emptyset \\
&R(\pi_1) = \{15\} \\
&R(\pi_2) = \{15,78\} \\
&R(\pi_3) = \{15,24,78\} \\
&R(\pi_4) = \{15,24,37,38,78\} \\
&R(\pi_5) = \{15,16,56,24,37,38,78\} \\
&R(\pi_6) = \{12,14,15,16,24,25,26,45,46,56,37,38,78\} \\
&R(\pi_7) = \{12,13,14,15,16,17,18,23,24,25,26,27, \\
&\qquad 28,34,35,36,37,38,45,46,47,48,56,57,58, \\
&\qquad 67,68, 78\}
\end{aligned}
\tag{34}
$$

On the other hand, by considering set theoretic difference, we obtain, according to above:

$$S(\pi_i) = F - R(\pi_i),\ 0 \le i \le 7 \tag{35}$$

'Significant' nodes and 'significant' levels of a classification tree

A decisive stage of the method consists of condensing the classification tree to the levels where a 'significant' node appears. This is based on an association criterion (or coefficient), statistically normalized, between a given partition π and an adequate structure retained from the resemblances between the elements of the set E to be classified.

For comparing two structures of a unique type, we are led to retain from the structure of the information concerning the resemblances only the related total pre-order on the set F of unordered element pairs from E (*ie* on the set of all subsets with two elements of E), called 'preordonance', associated with the similarity index S defined on E, as follows:

$$[\forall (p,q) \in F \times F];\ p \le q \longleftrightarrow S(p) \le S(q) \tag{36}$$

Thus the higher the rank of a given pair for the total preorder on F, the more similar are the components of the latter pair, according to S. Under these conditions, the preordonance $\omega(A)$ associated with the similarity established in table IV, is given by:

p : 58<18<57<13<35<17<34<23<45=47<67=68<26<14<46<36<37<16<25=27=28<48<56<24<38<12<78<15
r(p) : 1 2 3 4 5 6 7 8 9,5 11,5 13 14 15 16 17 18 20 22 23 24 25 26 27 28

(37)

where

$$\{r(p)/p \in F = P_2(A)\} \tag{38}$$

designates the 'average rank function' associated with the total pre-order. In this case, whatever the total and strict order compatible with the preceding total pre-order, $r(p)$ is the mean of the respective ranks, defined for the strict total order, of the different elements of the pre-order total class where p appears. This way the sum of all rank values (see (38)) is constant and equals $f(f+1)/2$, where $f = \text{card}(F)$.

By employing this ranking function we generalize in a suitable way a criterion initially conceived and computed in case where the structure of the resemblance information is a total and strict order on F (see (36)). In order to elaborate this criterion, a partition π on E is considered as defining a pre-order on F with two classes, $S(\pi)$ and R(π), defined above (see directly below relation (33)). According to the partition π, a separated pair (respectively a joined pair) is interpreted as composed of close (respectively distant) components. Then $S(\pi)$ precedes $R(\pi)$ for the quotient order; the partition π will be represented at the level of $F \times F$ by the cartesian product:

$$S(\pi) \times R(\pi) \tag{39}$$

whereas the total ordonance $\omega(E)$ is represented in the same manner by its graph:

$$gr[\omega(E)] = \{(p,q)/(p,q) \in F \times F \text{ and } p < q \text{ (strictly)}\} \tag{40}$$

Under these conditions, the association criterion between the ordonance $\omega(E)$ and the partition $\pi(E)$ that we denote by $C\ [\omega(E), \pi(E)]$ – can be put under the following set theoretic form

$$C(\omega,\pi) = \frac{s(\omega,\pi) - [r(\pi) \times s(\pi)/2]}{\sqrt{r(\pi) \times s(\pi)(f+1)/12}} \tag{41}$$

where

$$s(\omega,\pi) = \text{card}\{gr(\omega) \cap [S(\pi) \times R(\pi)]\} \tag{42}$$

and where $r(\pi) = \text{card}[R(\pi)]$ (respectively $s(\pi) = \text{card}\ [S(\pi)]$) with

$$f = \text{card}\ (F) = r(\pi) + s(\pi) \tag{43}$$

($f = 28$ in our example).

Whatever the case, $s(\omega, \pi)$ can be put under the following form:

$$s(\omega,\pi) = \sum \{\nu(q)/q \in R(\pi)\} \tag{44}$$

where $\upsilon(q)$ is the number of separated pairs, strictly on the left of q, for the preordonance ω.

We are going to directly use the formula (41) of the criterion in order to compare our preordonance

$\omega(A)$ (see (37)) with the sequence of the partitions $\{\pi_i/0 \leq i \leq 7\}$ determined above (see eqn (33)), by the hierarchical classification (see fig 8).

The use of $C(\omega,\pi)$ assumes implicitely a specific definition of the rank function and a statistical normalization procedure which holds in case when the preordonance $\omega(E)$ is fine enough (*ie* the number of tied ranks is small). Then, it is admissible to deal with our preordonance $\omega(A)$ (see eqn (37) above). The sequence of the values of (42), expressed by (44)

$$\{s\ (\omega, \pi_i)\ /0 \leq i \leq 7\} \tag{45}$$

is:

$$\begin{aligned}
&s\ (\omega, \pi_0) = \underline{0}\\
&s\ (\omega, \pi_1) = \underline{27}\\
&s\ (\omega, \pi_2) = 26 + 26 = \underline{52}\\
&s\ (\omega, \pi_3) = 25 + 23 + 25 = \underline{73}\\
&s\ (\omega, \pi_4) = 23+22+16+22+23= \underline{106}\\
&s\ (\omega, \pi_5) = 21+16+20+20+16+20+21 = \underline{134}\\
&s\ (\omega, \pi_6) = 15 + 11 + 15 + 12 + 15 + 12 + 11\\
&\qquad\qquad + 8 + 11 + 15 + 12 + 15 + 15 = \underline{167}\\
&s\ (\omega, \pi_7) = \underline{0}
\end{aligned} \tag{46}$$

For $s\ (\omega, \pi) = 0$, we have:

$$C\ (\omega, \pi) = -\{3r(\pi)s(\pi)\ /\ [r(\pi) + s(\pi) + 1]\}^{1/2} \tag{47}$$

By convention we will set $C\ (\omega, \pi) = 0$ if the product $r(\pi)\ \mathrm{x}\ (\pi)$ is null.

By introducing the increasing rate of $C\ (\omega, \pi)$ between two consecutive levels of the classification tree:

$$\tau(\omega, \pi_i) = C\ (\omega, \pi_i) - C\ (\omega, \pi_{i-1}) \tag{48}$$

$1 \leq i \leq 7$, one finally obtains, after calculation, table XII:

Table XII. Values of 'global' and 'local' level criteria.

i	$C(\omega, \pi_i)$	$\tau\ (\omega, \pi_i)$
0	0.000	
1	1.671	1.671
2	2.139	0.468
3	2.637	0.498
4	2.909	0.272
5	3.210	0.301
6	3.202	–0.008
7	0.000	–3.202

A 'significant' level does correspond to a local maximum of the distribution

$$\{C\ (\omega, \pi_i)\ /0 \leq i \leq 7\} \tag{49}$$

of the C criterion values, on the increasing sequence of the levels of the classification tree. It determines a partition which represents an equilibrium state in the automatic synthesis. In the above example there is only one 'significant' level, which is exactly the fifth. However, generally, there are different 'significant' levels, leading to several partitions giving, respectively, different pertinent degrees of data synthesis.

The most interesting consists of detection of 'significant' nodes. A 'significant' node does correspond to a local maximum of the distribution

$$\{\tau\ (\omega, \pi_i)\ /1 \leq i \leq 7\} \tag{50}$$

of (48) on the increasing sequence of the levels of the classification tree. For a given synthesis degree, a 'significant' node does indicate a completion stage of the underlying class, in the classification tree. In the above example (see fig 8), there are three 'significant' nodes marked by a star. Experience shows that a class which comprises 'significant' nodes, is clearly structured for understanding aspect. Let me indicate that in our most recent research [12, 13], in order to extract an interesting partition from a classification tree, we tended to forget the notion of a tree level and then, to retain a set of interesting nodes. The latter are either 'significant' or directly related by branching relation, according to the diagram tree, to 'significant' nodes.

Indices for interpretation: 'neutrality' degree of the elements of the set to be classified

One aspect of intrinsic validation with respect to a classification goal is relative to measure the neutral character of a given element x of E. As a matter of fact, there are elements which play an important role in leading the class forming. At the opposite, there are elements which are too 'neutral'; the position of some of the latter in their respective classes, may be difficult to interpret.

In this respect, the 'neutrality' degree of a given element α of the set to be classified (here A) is measured by the smallness of the observed variance $V\ (\alpha)$ of its proximities to the other elements of A. By designating p the cardinality of A, we have:

$$P(\alpha) = \frac{1}{p-1} \sum \{[Q(\alpha, b) - Q(\alpha)]^2 / b \in A - \{\alpha\}\}$$

where

$$Q(\alpha) = \frac{1}{p-1} \sum \{Q(\alpha, b) / b \in A - \{\alpha\}\} \tag{51}$$

Table XIII, illustrates the behaviour of the defined statistic V in the framework of our example, on the

basis of the table V. The last column gives the respective ranks of the different elements of A, according to decreasing values of V.

Table XIII. Values of the variance V and the associated ranks.

j	$V(a^j)$	$r(a^j)$
1	0.8415	3
2	0.3414	7
3	0.3556	6
4	0.6037	4
5	0.9522	2
6	0.1483	8
7	0.4105	5
8	1.0109	1

Let us indicate that the analysis of the variance of proximities between the elements of the set to be organized, leads to a rich family of 'seriation' and 'classification' methods [14, 15].

How to reach a discriminant probability scale, when the object set becomes large

As a matter of fact, it can be established that, when n = card (O) tends to infinity, the probability value given by equation (21), tends 'very quickly' to zero value (respectively, one value), in the case where the concerned link is negative (respectively positive). For this behaviour, the influence of the link intensity remains weak.

The indices as those established in association table (19) by means of the equation (20), are called 'local' indices; because the comparison between two given attributes is logically independent form the context of the other attributes belonging to A.

Consider the substitution of the 'local' indices

$$\{Q_l(j,k)/1 \leq j < k \leq p = \mathrm{card}(A)\} \tag{52}$$

by 'global' indices

$$\{Q_g(j,k)/1 \leq j < k \leq p\} \tag{53}$$

where $Q_g(j, k)$ is obtained from $Q_l(j, k)$ by standardization, $1 \leq j \leq k \leq p$. More precisely,

$$Q_g(j,k) = \frac{Q_l(j,k) - \mathrm{moy}_e(Q)}{\sqrt{\mathrm{var}_e(Q)}} \tag{54}$$

where $\mathrm{moy}_e(Q)$ and $\mathrm{var}_e(Q)$ are respectively the empirical mean and variance of the distribution (52) of the Q_l-values.

A minute statistical analysis [16, 17] enables to assess, under an independence hypothesis, as figured in (23), the reference to the probabilistic index

$$P(a^j, a^k) = \phi[Q_g(a^j, a^k)] \tag{55}$$

$1 \leq j \leq k \leq p$, where ϕ denotes the normal cumulative distribution function.

By this transformation the probability scale becomes discriminant enough in order to evaluate in a relative manner the observed relations between the descriptive attributes of the object set O. Considerable work has been devoted to extend and to assess this approach for most general types of data. These types will be globally expressed in the next chapter.

Now, we are going to have in the simplest situation and on the basis of the previous example, the general outline of the elaboration method of a probabilistic similarity index on the set O of objects (respectively C of concepts).

Probabilistic similarity index on the object set O, described by boolean attributes (case of the example of table I)

The first step consists of defining for j, $1 \leq j \leq p$ ($p = 8$ in our example), respectively, a 'raw' similarity index s_j (o_i, $o_{i'}$) between the two objects o_i and $o_{i'}$, for comparison.

Proposition of a similarity index s_j

Substitute to the incidence table I:

$$\{\varepsilon_i^j / 1 \leq i \leq n, 1 \leq j \leq p\} \tag{56}$$

the table

$$\{\eta_i^j / 1 \leq i \leq n,\ 1 \leq j \leq p\} \tag{57}$$

where

$$\eta_i^j \frac{\varepsilon_i^j}{\sqrt{\sum_{1 \leq k \leq p} \varepsilon_i^j}} \tag{58}$$

then, we set

$$s_j(o_i, o_{i'}) = \frac{1}{p} - \frac{1}{2}(\eta_i^j - \eta_i^{i'})^2 \tag{59}$$

Thus, for example,

$$s_1(o_1, o_2) = \frac{1}{8} - \frac{1}{2}\left(\frac{0}{\sqrt{4}} - \frac{0}{\sqrt{3}}\right)^2 = 0.125$$

$$s_4(o_1, o_2) = \frac{1}{8} - \frac{1}{2}\left(\frac{1}{\sqrt{4}} - \frac{1}{\sqrt{3}}\right)^2 = 0.122$$

$$s_7(o_1, o_2) = \frac{1}{8} - \frac{1}{2}\left(\frac{0}{\sqrt{4}} - \frac{1}{\sqrt{3}}\right)^2 = -0.042$$

$$s_3(o_1, o_2) = \frac{1}{8} - \frac{1}{2}\left(\frac{1}{\sqrt{4}} - \frac{0}{\sqrt{3}}\right)^2 = 0.000$$

Notice that in general the mathematical and logical structure of the data table has to intervene closely in the conception of the raw similarity index s_j (o_i, $o_{i'}$) [respectively s_j (C_i, $C_{i'}$)] between two given objects (respectively concepts) (see fig 5).

In the second step, the raw index s_j (o_i, $o_{i'}$) is normalized with respect to the empirical distribution

$$\{s_j(o_i, o_{i'})/1 \leq i, i' \leq n\} \tag{60}$$

which comprises n^2 observed values, in order to obtain

$$S_j(o_i, o_{i'}) = \frac{s_j(o_i, o_{i'} - m_e(s_i)}{\sqrt{\text{var}_e(s_j)}} \tag{61}$$

where m_e (s_j) and var_e (s_j) are respectively the empirical mean and variance of the preceding distribution (60).

As mentioned, the empirical mean and variance are relative to n^2 values of s_j (o_i, $o_{i'}$). In our example, as in many real cases, theses values are equally weighted. But this condition does not hold if the weight of an object cannot be considered constant when the latter object varies in the object set O. Indeed, when we have to deal with classifying a set C of concepts, the weight of a concept c is not constant when c varies from C.

s_j (o_i, $o_{i'}$) is called 'normalized contribution of the j^{th} descriptive variable in comparison of objects o_i and $o_{i'}$'.

The third step

$$S(o_i, o_{i'}) = \sum_{1 \leq j \leq m} S_j(o_i, o_{i'}) \tag{62}$$

is the sum of the normalized contributions of the descriptive variables.

The fourth step consists of the global normalization of the preceding indices, with respect to the following empirical distribution

$$\{S(x, y)/\{x, y\} \in P_2(O)[\text{respectively } P_2(C)]\} \tag{63}$$

where $P_2(O)$ (respectively $P_2(C)$, denotes the set of unordered element pairs of O (respectively C). Then we obtain from S (o_i, $o_{i'}$)

$$Q_g(o_i, o_{i'}) = \frac{S(o_i, o_{i'}) - m_e(S)}{\sqrt{\text{var}_e(S)}} \tag{64}$$

where $m_e(S)$ and $\text{var}_e(S)$ are respectively the mean and the variance of the preceding distribution (63), $1 \leq i \leq i' \leq n$.

The final step consists of computing the probabilistic index

$$P(o_i, o_i') = Pr\{Q_g(x*, y*) \leq Q_g(o_i, o_i')\} \tag{65}$$

This probability is computed under independence hypothesis, as figured in (23). If the number p of variables is large enough, the following approximation is very accurate:

$$P(o_i, o_i') = \phi[Q_g(o_i, o_i')] \tag{66}$$

$1 \leq i < i' \leq n$, where ϕ indicates the $N(0, 1)$ normal cumulative distribution function.

By taking into account the extension of this index for all types of data (see next section), the reason of the diagram of figure 6 becomes clear.

Let us now show the output of the computer program CHAVL (Classification Hiérarchique par AVL) on the set of 10 objects given by the treated example (see table I). The formulas (59), (61), (62), (64) and (66) provide the calculation of the similarity table. To mention is that this program has been set up by P Peter (Ireste, Nantes) in collaboration with H Leredde (university Paris-Nord) and with IC Lerman (university Rennes 1-Irisa).

In the following output, the table of the dispersions of the respective elements of the set to be classified, is associated with table XIII. On the other hand, the table of the 'global' and 'local' level statistics corresponds to table XII.

Hierarchical classification program CHAVL

Study title: Incidence table test

SIMOB stage

Recall parameters

number of objects 10
number of variables 8
presence-absence attributes
data lay-out (3 x 8 *i* 2)

End of SIMOB stage

AVLR stage

Polish tree representation

–9 –7 1 –6 –3 –1 2 8 4 10 –8 –5 –4 3 –2 6
9 7 5 0

End of AVLR stage

INTRP stage

Ranking of the data units by decreasing values of dispersion index

Element 10:o 10 dispersion: 0.36271
Element 7:o 7 dispersion: 0.43063
Element 1:o 1 dispersion: 0.45478
Element 5:o 5 dispersion: 0.62854
Element 9:o 9 dispersion: 1.04031
Element 6:o 6 dispersion: 1.06048
Element 2:o 2 dispersion: 1.09864
Element 3:o 3 dispersion: 1.10182
Element 8:o 8 dispersion: 1.11017
Element 4:o 4 dispersion: 1.42690

Level statistics

	Level	Global statistics	Local statistics
	1	1.8040	1.8040
	2	2.4660	0.6019
1 maximum	3	3.2918	0.8258
	4	3.8923	0.6005
	5	4.02541	0.1317
2 maximum	6	4.3176	0.2935
	7	4.5721	0.2546
3 maximum	8	4.8333	0.2612
	9	–0.1124	–4.9457

End of INTRP stage

DESAB stage
Parameters of the stage

Number of elements	10
Number of represented levels	9
One level by 1 is represented	9
Length of a line	80

End of DESAB stage

End of CHAVL job

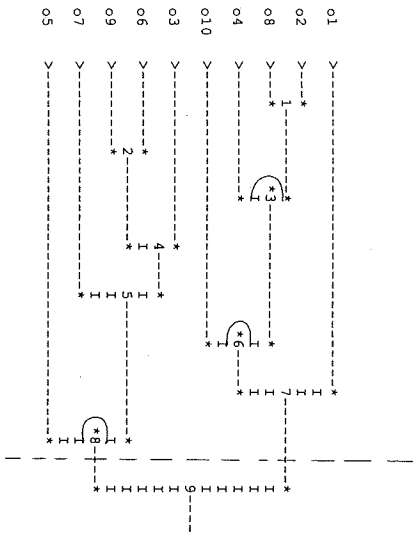


Fig 9. Classification tree on the object set.

Crossing between a partition on the object set and a partition on the attribute set

As mentioned just above the second section, we are going to illustrate in the simplest case the management of ARBA with respect to ARBO. For this purpose, the respective partitions associated with the most 'significant' levels of ARBA (see fig 8) and ARBO (see fig 9) were crossed.

Table XIV is obtained from the data table I, by reordering its rows and columns, according to the couple of partitions described above. The former is into two classes $\{C_1, C_2\}$ on O; the latter is into three classes $\{A_1, A_2, A_3\}$ on A.

Table XIV. Classificatory rankings of rows and columns.

		A_1			A_2		A_3		
	A	a^1	a^5	a^6	a^2	a^4	a^3	a^7	a^8
	O								
C_1	o_2	0	0	0	0	1	0	1	1
	o_3	0	0	0	0	1	1	1	1
	o_4	0	0	0	0	0	1	1	1
	o_{10}	1	0	0	1	1	0	1	1
	o_1	0	0	1	0	1	1	0	1
C_2	o_6	1	1	0	0	1	0	0	0
	o_9	1	1	1	0	1	0	0	0
	o_3	1	1	0	1	1	0	0	0
	o_7	1	1	1	0	0	0	1	0
	o_5	1	1	0	0	0	1	0	0

Then the association coefficients which enable to compare an object class with an attribute class are introduced. The interest of these coefficients grows when the sets O and A become large. The presented coefficients are more general, since they concern the comparison of two disjoint attribute sets.

Some definitions are necessary. Let me denote here by A a class of boolean attributes. In this application A will become A_1, A_2 or A_3 (see above). A will be called a class of 'oriented' boolean attributes if the belonging of an attribute a to A excludes the belonging of the associated opposite attribute ã to A $[(\forall x \in O), \tilde{a}(x) = 1 - a(x)]$.

Relative to a given class of 'oriented' boolean attributes denoted by A, we introduce the associated class $\tilde{A}$, composed of the respective complemented attributes; namely

$$\tilde{A} = \{\tilde{a}/a \in A\} \qquad (67)$$

For a given attribute a, describing one aspect of the object set O, we classically have

$$a(x) = 1 \text{ or } 0[\text{respectively } \tilde{a}(x) = 0 \text{ or } 1] \quad (68)$$

according to presence or absence of the attribute a on the object belonging to O.

We obviously have

$$(\forall x \in O),\ a(x) + \tilde{a}(x) = 1 \quad (69)$$

Now the following function is introduced

$$A(x) = \frac{1}{c(A)} \sum_{a \in A} a(x) \quad (70)$$

where $c(A)$ designates the cardinality of A. The ratio (70) represents the proportion of attributes of A of which the value is 'true' on the given object x.

Now there is an analogous relation to (69):

$$(\forall x \in O),\ A(x) + \tilde{A}(x) = 1 \quad (71)$$

Consider now two attribute classes A and B, made up each of 'oriented' boolean attributes. On the other hand, assume, as usual for the aim, that A and B are disjoint ($A \cap B = \emptyset$: there is no attribute which is common to A and to B). Then, define the 'raw' index:

$$\sigma(A,B) = \nu(A \cap B) = \sum_{x \in O} A(x)B(x) \quad (72)$$

$\sigma(A, B)$ represents exactly the adequate generalization of $s(a, b)$ (see [9]) defined in order to associate two boolean attributes.

Statistical normalization (centering and reduction) of $\sigma(A, B)$ with respect to probabilistic hypothesis of no relation [2], leads to the multiplicative factor $\sqrt{n-1}$, – to the coefficient:

$$\chi(A,B) = \left[\sum_{(a,b) \in A \times B} (p_{a,b} - p_a p_b)\right] \Big/ \left\{\left[\sum_{A \times A} (p_{a,a'} - p_a p_{a'})\right]\left[\sum_{B \times B} (p_{b,b'} - p_b p_{b'})\right]\right\}^{1/2} \quad (73)$$

We have

$$-1 \leq \chi(A,B) \leq +1 \quad (74)$$

Otherwise, if A^* (respectively B^*) is a class of independent random attributes, associated with A (respectively B) and if A^* and B^* are independent, then the limit distribution of $\sqrt{n}\,\chi$ (A^*, B^*) is the standardized normal distribution.

In our case (see above) a given class C_1 or C_2 defines a boolean attribute. One may denote by c^1 (respectively c^2) the boolean attribute represented by the subset C_1 (respectively C_2). Then, the following table of the above defined coefficients is obtained

$$\{\chi(A_i, c^j)/1 \leq i \leq 3, j = 1,2\} \quad (75)$$

Table XV. Normalized coefficients between attribute and object classes.

	A_1	A_2	A_3
C_1	–0.833	0.143	0.898
C_2	+0.833	–0.143	–0.898

This table shows the direction of the global associations between $\{A_1, A_2, A_3\}$ on the one hand and $\{c^1, c^2\}$ on the other. Since c^1 and c^2 are complementary with respect to the set O, χ (A_i, c^1) and χ (A_i, c^2) are opposite, $1 \leq i \leq 3$.

Methodological extensions

The combinatorial and statistical approach in data analysis is now recognised as a very rich methodological domain. The valuable and impressing synthetic review of Arabie and Hubert [4] is not complete in covering the whole field. Nevertheless, this paper mentions about 500 references, including many books, mostly for the last ten years. For our part, we are going to express those aspects where we have brought, notably in collaboration with other researchers, significant contribution. We will also indicate methods where our approach of probabilistic similarity can be integrated. Therefore, our references are rather concentrated; however, more general bibliography can be consulted by considering the respective references of the cited publications.

The above-mentioned methodological aspects can directly be of interest for data provided by biological sequences.

General types of data

In the framework of the likelihood linkage analysis (LLA) method and in fact, very generally, one can distinguish two fundamental types of data. The first is given by the 'relational system of Tarski' [18]:

$$T = < O; R_1, R_2, \ldots, R_j, \ldots, R_p > \quad (76)$$

where $R_1, R_2, \ldots, R_j, \ldots, R_p$ are p relations defined on the set O of elementary objects. In the case described here, the relation R_j is defined by the j^{th} descriptive attribute (or 'variable') a^j, $1 \leq j \leq p$. Under these conditions, the relations R_j, $1 \leq j \leq p$, are supposed to have (or to be reduced to) the same combinatoric type and then, to be of same arity. Thus, for example, in the

classical cases where the variables aj, $1 \leq j \leq p$, are either nominal or ordinal variables, the associated relations R_j, $1 \leq j \leq p$, are, respectively, partitions or total pre-orders of O. In these cases, the common arity q of the respective relations is equal to 2. Indeed, with some generality, in the classical quantitative or qualitative data analysis, it is sufficient to consider $q = 1$, 2 or 4. A q – ary relation R_j is represented by a structured subset of O^q. Valued relations are also considered and represented by weighted structured subsets of O^q, if q is the arity of the relations concerned. In our small example, the respective relations determined by boolean attributes are unary ($q = 1$) and not valued (see fig 5).

The second fundamental type of data can be figured by the system

$$S = < C; R_1, R_2, \ldots, R_j, \ldots, R_p > \qquad (77)$$

where C is a set of concepts or classes. Herein the data correspond to statistical distributions of each R_j, $1 \leq j \leq p$, on each concept (or class) c belonging to C.

We may note that the data structure concerned by corrrespondence analysis (contingency table or even horizontal juxtaposition of contingency tables) is a particular case of the system S. For the latter, each R_j, $1 \leq j \leq p$, is defined by a partition relation, associated with a nominal qualitative variable.

For both systems T and S we apply the general scheme described in the final part of the first section of this article.

For references concerning work done in our research environment, in order to realize the preceding scheme for T and for S see [9, 17, 19–26].

Association coefficients and proposition of a dissimilarity index

In the conception of an association coefficient between two relational variables, according to the LLA, the centered index is reduced by means of the standard deviation of the random raw index (see equation (20)). There is another point of view which consists of reducing the centered index by means of the maximal value that it can reach, under the constraints determined by the hypothesis of no link (h.n.l.). This approach leads to very difficult problems of combinatorial optimization [26].

The association coefficients between relational variables considered above, have a 'total' nature. But one may wish, in the organization of the relationships between the descriptive variables (construction of ARBA in the scheme ending section 1), to neutralize the influence of an external group of variables. For this purpose, theoretical studies have been performed for elaborating partial association coefficients [27, 28]. On the other hand, very interesting experiments have been led by A SBII (unpublished thesis, 1988).

We have already emphasized that whatever the nature of the set E to be organized (see diagram of fig 6), we end with the resemblance matrix:

$$\{P(x, y) / \{x, y\} \in P_2(E)\} \qquad (78)$$

of probabilistic indices of similarity between the elements of E. However, many methods in data representation assume a given dissimilarity or distance index on E. It is always possible to deduce a distance index from a dissimilarity index by adding a minimal constant. The dissimilarity index which can be the most naturally associated with the probabilistic index $P(x, y)$, is the amount of information of the event of which the probability is precisely $P(x, y)$. Under these conditions, we substitute for the matrix (78), the following

$$\{-\log_2[P\{x, y)] / \{x, y\} \in P_2(E)\} \qquad (79)$$

in order to apply a data representation method based on a dissimilarity matrix.

Other types of data representation structures

By ordering these structures, according to their metrical nature, one can mention among the most used: i) the total or partial orders or preorders; ii) the additive trees; and iii) the factorial analyses of distance tables (multidimensional scaling).

We just expressed direction in order to employ our similarity indices (see (79)) in cases ii) [29] and iii).

Concerning i) above, let us consider, for reasons of simplicity, the case where the attributes are boolean (presence-absence variables). 'Seriation' problems appeared in archeological methodology for this latter data structure. The matter is to seek for a couple of two corrresponding rankings, eventually with ties, on respectively the object set O and the attribute set A. Then a couple of permutations on rows and columns of the incidence data table associated with description are to be found out, such that a diagonal form appears with high density of 1 values. We have already mentioned our contribution to this problem (see references expressed following table XIII). We may also cite French work [30, 31].

Partial orders or pre-orders have appeared as representation structures in the context of data providing in the didactic domain. The purpose is building 'implication' graphs between stimuli or classes of stimuli. In this respect, the symmetrical notion of probabilistic similarity has to be replaced by an oriented notion of probabilistic implication. The obtained index enables to evaluate the magnitude of the degree with which a positive value of an attribute a, implies a positive value of an attribute b [32, 33].

Algorithmic problems set up by synthetic organization of 'large' data sets

The synthetic organization may either correspond to clustering scheme or any other representation structure (see previous paragraph). Here, we are mainly concerned with ascendant hierarchical classification. In this framework, the last ten years, new ideas have arisen and have been experimented with. These ideas use principally notions called: i) 'reciprocal nearest neighbours' and 'reducible neighbourhoods' [34]; and ii) 'parallel hierarchical classification' [35, 36].

Recently we proposed [37] a thesis subject where the preceding algorithmic ideas were combined in an optimal way. The goal to reach is classification of 'very large' data sets of which the size can attain several 10 000 of elements. Indeed, the molecular biology field is concerned with this size problem.

Statistical validity and stability of a classification

Let us designate by E the set to be classified, and, according to the notations of the diagram which ends the first section, if $E = O$ (respectively C), we set $E^* = A$; dually, if $E = A$, we set $E^* = O$ (respectively C). Under these conditions, the problems of statistical validity and stability can be situated at two levels: i) relatively to the size of E^*; and ii) relatively to adding (or extracting) elements from E (respectively E^*).

Concerning i) we have mainly studied the case where E is defined by the attribute set A [36]. However, it is also of importance to study this same problem in case where E correponds to the set O (respectively C) of described elementary objects (respectively concepts).

Some directions of work

One of the general directions of work that we hope to lead concerns interaction between the above described general approach and most problems set up in the context of typological organization of amino-acid sequences. We will begin by following the statistical expression of the similarity information as considered by the biologist (*eg* Dayhoff matrix). However, to the numerical structure will be added a combinatorial structure having an ordinal nature. The latter is assumed to be more robust and more synthetic than a numerical one. Thus, for example, from the Dayhoff matrix we only retain ranking (with ties) on the set of all pairs of amino-acids, including comparison between each amino-acid and itself. In order to perform pairwise comparison between all the sequences, after multiple alignment, estimations are proposed to respectively associate a recognized amino-acid with deletion on the one hand and with an unknown amino-acid on the other. The first experiments, led in collaboration with P Peter (Ireste Nantes) (unpublished work), are very encouraging. They address two data sets; the former comprises 68 aligned cytochrome sequences and the latter 42 aligned globin sequences. These data have been available thanks to JL Risler (Centre de Génétique Moléculaire du CNRS).

The second stage of our approach consists of studying the intrinsic nature of the similarity information that the biologist wishes to induce from data. The big problem is defining 'relevant' description. By considering symbolic characterizations provided by Machine Learning techniques, the latter description must integrate formal operations on a family of chains of graphic symbols, where each symbol corresponds to the representation of an amino-acid. This family can for example be defined by all chains formed by consecutive symbols and having a given lenght. Thus, the similarity concept has to take into account the preceding operations. For example, some invariance properties for the similarity value, under a given class of transformations, can be required. Finally, the result of the cluster analysis includes a logical parameter 'value' which corresponds to the allowable family of operations. This direction of work is of importance, specially by taking into account statistical significance aspects in the spirit of the LLA methodology.

Whatever the retained descriptive structure and the adopted characterization method are, the problem of objective evaluation of the results, of computational and statistical data analysis, in terms of new biological knowledge, remains crucial.

References

1 Saurin W (1991) Comparaison de séquences biologiques. *Cahier IMABIO 2*, 'Informatique et Génomes. Traitement de l'information des séquences biologiques', CNRS, 49–53

2 Lerman IC (1981) *Classification et analyse ordinale des données*. Dunod, Paris

3 Lerman IC (1991) Foundations of the likelihood linkage analysis (LLA) classification method. *Applied Stochastic Models and Data Analysis*, John Wiley, Vol 7, 69–76

4 Arabie P, Hubert LJ (1992) Combinatorial data analysis *Annu Rev Psychol* 43, 169–203

5 Fisher D, Langley P (1985) Approaches to conceptual clustering. *In: Proceedings of IJCAI 85*, Los Angeles, 691–697

6 Quinqueton J (1991) Méthodes d'apprentissage sur des descriptions attributs/valeurs: CALM. *Cahier IMABIO 2*, 'Informatique et Génomes. Traitement de l'information des séquences biologiques', CNRS, 23–29

7 Lafaye JY (1979) Une méthode de discrétisation de variables continues. *Revue de statistique appliquée*, n° 2

8 Lerman IC, Peter P (1989) Classification of concepts described by taxonomic preordonnance variables with multiple choice. Application to the structuration of a species

set of phlebotomine. *In: Data Analysis, Learning Symbolic and Numeric Knowledge*, Proceed of the Conf Antibes, sept 11–14, (Diday E, ed) Inria, Nova Science Publishers Inc, New York, 73–86

9 Ouali-Allah M (1991) *Analyse en préordonnances des données qualitatives. Applications au données numériques et symboliques*, Thèse de doctorat, université de Rennes I

10 Jaccard P (1908) Nouvelles recherches sur la distribution florale. *Bull Soc Vaud Sci Nat* 44, 223–2790

11 Nicolaü F (1980) *Critérios de analise classificatoria hiérarquica baseados na furçao de distribuiçao*. Thèse de Ph. D faculté des sciences de Lisbonne

12 Lerman IC, Ghazzali N (1991) What do we retain from a classification tree? An experiment in image coding. *In: Symbolic-Numeric Data Analysis and Learning*. Proceed of the conf Versailles, sept 18–20 (Diday E, Lechevallier Y, eds) Inria, Nova Science Publishers Inc, New York, 27–42

13 Ghazzali N (1992) *Comparaison et réduction d'arbres de classification, en relation avec des problèmes de quantification en imagerie numérique*. Thèse de doctorat, université de Rennes I

14 Lerman IC (1972) Analyse du phénomène de la 'sériation'. *Rev Math Sci Hum* 38, 39–57

15 Leredde H (1979) *La méthode des pôles d'attraction; la méthode des pôles d'agrégation: deux nouvelles familles d'algorithmes en classification automatique et sériation*. Thèse de 3e cycle, université Paris VI

16 Lerman IC (1984) Justification et validité statistique d'une échelle [0, 1] de fréquence mathématique pour une structure de proximité sur un ensemble de variables observées. *Publ Inst Stat Univ Paris* 29, 27–57

17 Daude F (1992) *Analyse et justification de la notion de ressemblance entre variables qualitatives dans l'optique de la classification hiérarchique par AVL*. Thèse de doctorat, université de Rennes I

18 Tarski A (1954) Contribution to the theory of models, I et II. *Indagationes Mathematicae* 16, 572–588

19 Lerman IC (1983) Association entre variables qualitatives ordinales nettes ou floues. *Stat et An des Don* 8, 41–73

20 Lerman IC (1984) Analyse classificatoire d'une correspondance multiple, typologie et régression. *In: Data Analysis and Informatics*, III (Diday E *et al*, eds) North Holland, 193–221

21 Lerman IC (1987) Construction d'un indice de similarité entre objects décrits par des variables d'un type quelconque. Application au problème du consensus en classification. *Rev Stat Appl* 35, 39–60

22 Lerman IC (1992) Conception et analyse de la forme limite d'une famille de coefficients statistiques d'association entre variables relationnelles, I et II. *Rev Math Inf Sci Hum* 118, 35–37 et 119, 75–100

23 Peter P (1987) *Méthodes de classification hiérarchique et problèmes de structuration et de recherche d'informations assistées par ordinateur*. Thèse de doctorat, université de Rennes I

24 Tallur B (1988) *Contribution à l'analyse exploratoire de tableaux de contingence par la classification*. Thèse de doctorat ès sciences, université de Rennes I

25 Lerman IC (1987b) Maximisation de l'association entre deux variables qualitatives ordinales. *Rev Math Sci Hum* 100, 49–56

26 Lerman IC, Peter P (1988) Structure maximale pour la somme des carrés d'une contingence aux marges fixées; une solution algorithmique programmée. *Rairo/Oper Res* 22, 83–136

27 Lerman IC (1983) Indices d'association partielle entre variables qualitatives nominales. *Rairo/Oper Res* 17, 213–259

28 Lerman IC (1983) Indices d'association partielle entre variables qualitatives ordinales. *Publ Inst Stat Univ Paris* 28, 7–46

29 Barthelemy JP, Guenoche A (1988) *Les arbres et les représentations des proximités*, Masson, Paris

30 Guenoche A (1987) Méthodes combinatoires de sériation à partir d'une dissimilarité. *Cinquièmes Journées Internationales 'Analyse des Données et Informatique'*, 29 sept–20 oct, Versailles, France, Inria, North Holland

31 Marcotorchino F. (1991) Seriation problems: an overview. *Applied Stochastic Models and Data Analysis*, John Wiley, 7, 139–151

32 Lerman IC, Gras R, Rostam H (1981) Elaboration et évaluation d'un indice d'implication pour des données binaires, I et II. *Rev Mat Sci Hum* 74, 5–35, et 75, 5–47

33 Gras R, Larher A (1992) L'implication statistique, nouvelle méthode d'analyse des données. *Rev Math, Inf Sci Hum,* 120

34 Bruynooghe M (1989) *Nouveaux algorithmes en classification automatique applicables aux très gros ensembles de données, rencontrés en traitement d'images et en reconnaissance des formes*. Thèse de doctorat d'état, Université de Paris VI

35 Lerman IC, Peter P (1984) *Analyse d'un algorithme de classification hiérarchique 'en parallèle' pour le traitement de gros ensembles*. Rapport de recherche n° 339, Inria, Le Chesnay

36 Lerman IC (1986) Rôle de l'inférence statistique dans une approche de l'analyse classificatoire des données. *Journal de la Société de Statistique de Paris* 4, 238–252

37 Lerman IC *et al* (1991) *Votre thèse à l'Irisa*, Irisa (Inria-CNRS)