APPLIED STOCHASTIC MODELS AND DATA ANALYSIS, VOL. 7, 63–76 (1991)

FOUNDATIONS OF THE LIKELIHOOD LINKAGE ANALYSIS (LLA) CLASSIFICATION METHOD

I. C. LERMAN
IRISA, Campus de Beaulieu, Avenue du Général Leclerc, 35042 Rennes cédex, France

SUMMARY

The aim of this paper is to present the concepts underlying an approach to data analysis using a hierarchical classification. The data can be provided by observation, experiment or knowledge. We begin by presenting the classical view of the context of data representation, in which the algorithm of hierarchical ascendant construction of the classification tree is set. The main notion in our method is one of 'similarity'. The latter must be elaborated in the best way, taking into account the mathematical nature of the objects to be compared. In this elaboration, we adopt a set theoretic and combinatoric representation of the descriptive attributes, which are interpreted in terms of relations. On the other hand, we introduce a probability scale for similarity measurement by using a likelihood concept.

KEY WORDS Hierarchical classification Relational attributes Probabilistic association coefficients

1. INTRODUCTION

Likelihood linkage analysis (LLA) is a methodology for grouping data into 'significant' classes and subclasses, using an algorithm of hierarchical classification. The latter is built in an ascendant way be successive agglomerations. However, this approach has many more than 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. The data can be provided by observation, experiment or knowledge. Additionally, this method introduces a most original notion of 'statistics' for measuring statistical relationships and proximities, namely, the 'likelihood' concept. Thus, we set up the 'likelihood' notion as part of the 'resemblance' notion. This principle also underlies the 'information theory' formalism, in which the higher the amount of information quantity, the more unlikely is the event concerned. In our case the events correspond to the observed relations.

The flexibility of the method enables us to take into account any combinatorial and logical structure of which the modality set of a given descriptive 'variable' (we also say 'attribute' or 'parameter') is provided. This structure is induced by the type of attribute, and in this respect we distinguish the following types: 'presence–absence', 'modality', (category) of a 'qualitative attribute', 'nominal qualitative', 'ordinal qualitative', 'taxonomic', 'preordonnance', 'taxonomic preordonnance', 'numerical' 'weighted graph', etc. Some of these types (taxonomic, preordonnance, taxonomic preordonnance) have been very recently introduced to treat 'data' faithfully where the value set of a given descriptive attribute is structured to a large extent by expert knowledge of the field studied.

8755–0024/91/010063–14$07.00
Received 20 November 1990

We show that even in the case of artificial intelligence data, all the relevancy of a double entry data table holds. Such a table cross refers a set U of 'data units' and a set A of 'descriptive attributes'. Most often U is either a set O of elementary objects or a set C of disjoint object classes. In the latter case, each object class represents a concept. One may not be able to observe A on a sample of objects representing a given concept, but only to have at one's disposal expert knowledge concerning the concept. In this case, the value of a given attribute associated with a potential object or with a concept (intentionally defined), may be a logical formula on the modality set of the attribute.

Regarding a data table indexed by the cross product $U \times A$, the LLA method handles two complementary and dual problems. The former is of 'understanding' and the latter of 'management'. The tree organization of A into 'significant' classes and subclasses of statistical association on U contributes in a relevant way to the first problem. On the other hand, the same type of organization of the set U into 'significant' proximity classes and subclasses contributes in an effective way to the second problem. The 'significant' nodes correspond to different synthesis levels in the class completion. A given class underlying a significant node can be considered as a principal component of a larger class containing it and ending with a 'significant' node.

We have built around the classification method a set of association measure coefficients relating to the cross referencing between a given attribute or a class of attributes and a class of elementary objects or a class of concepts. This association may also be considered to be an organized system relating either elementary object classes or concept classes.

2. 'CLASSICAL' VIEW OF ASCENDANT HIERARCHICAL CLASSIFICATION (AHC)

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:

$$(\mathcal{O}, \mu_{\mathcal{O}}, d) \tag{1}$$

where $\mathcal{O}$ is a finite set of objects, provided by a positive measure $\mu_{\mathcal{O}}$ and where d is a dissimilarity or distance index defined on $\mathcal{O}$.[1-8]

From (1) we may deduce the following table for the distances between weighted objects:

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

where $P_2(\mathcal{O})$ is the set of unordered object pairs.

One extends the notion of distance (or dissimilarity) d, between elements of $\mathcal{O}$, to a notion of distance (or dissimilarity) δ between subsets of $\mathcal{O}$. Thus to the triplet (1) we associate the following triplet:

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

the set of all

where P is subsets of $\mathcal{O}$ and where μ_P is a positive measure on P, deduced from $\mu_{\mathcal{O}}$.

Let $\mathbb{R}_+$ denote the set of real positive numbers. The distance or dissimilarity δ can be put in the following mapping form:

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

where

$$\left.\begin{array}{c}[\forall (X,Y) \in P \times P] \\ \delta(X,Y) = f[\{ d(x,y) / (x,y) \in (X \cup Y) \times (X \cup Y)\},\ \mu_{X \cup Y}]\end{array}\right\} \quad (5)$$

where the function f is to be defined and where $\mu_{X \cup Y}$ is the restriction of $\mu_{\mathcal{O}}$ on $X \cup Y$:

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

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

The algorithm of ascendant hierarchical classification, using δ to build a classification tree on $\mathcal{O}$, is a 'trivial' mathematical principle: 'At each step, join the class pairs which realize the minimum value of δ'.

Example

Let $\mathcal{O} = \{1, 2, 3, 4, 5, 6,\}$ be the object set, in which each integer code denotes a single object. A classification tree $A(\mathcal{O})$ on $\mathcal{O}$ may have the form shown in Figure 1.

However, this mathematical 'triviality' does not entail computational and statistical 'trivialities'.

In these terms, we expand below different problems concerning these two last aspects. Some of them—of a theoretical nature—remain very difficult to solve. We end with some general remarks concerning the 'classical' view, which justify, and demonstrate the relevance of, our approach.

Problem 1: Table management of the δ indices between the classes built up

We may compare the hierarchical ascendant construction of a classification tree to the 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 expression (3), above). Then, it is of importance from a computational point of view to have a formula—called the 'reactualization formula'—of the following form:

$$(T_{K+1}, \mu_{K+1}) = \varphi(T_K, \mu_K) \quad (7)$$

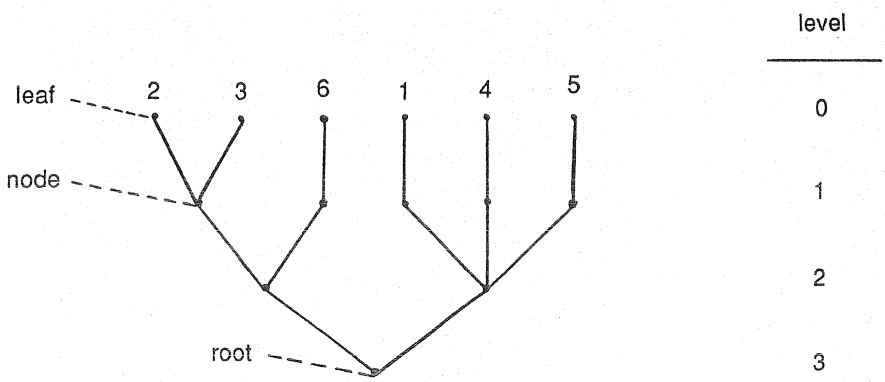

Figure 1. $A(\mathcal{O})$

which gives the system state at the level $(K+1)$ after joining the nearest classes of the Kth level, determined on the basis of $T_K \cdot T_0$ is the table of indices between singleton classes, provided by (2).

Until recently all the transformations φ were given for the union of only a single pair of classes.[6,9,10]

However, there is a real problem of aggregation in cases where several class pairs realize 'at the same time' (i.e. at the same tree level) the minimum value of δ. This circumstance occurs in particular because on one hand the descriptive scales of $\mathcal{O}$—leading to the definition of the distance d—are poor, and on the other the size [card $(\mathcal{O})$] of the object set $\mathcal{O}$ is 'large'. Then we have proposed for the most classical indices δ reactualization formulas for cases of multiples aggregations.[11]

In addition to (7) we have to suppose an informatic representation of a classification tree, to collect the sequence of the aggregations and to end at the graphic representation. For our own, we adopt a polish representation of the classification tree.

Problem 2: Hierarchical classification of 'large' data sets

The notion of a 'large' data set, has a relative meaning. The correct expression of the problem is relative to the definition of algorithms which decrease—in time and space—the computational complexity.

An important notion introduced by M. Bruynooghe in 1977[3] is the 'reducibility' property of the aggregation criterion δ [cf. (5)], according to which, for a fixed radius and in relation to an exhaustive system of classes, the ball of which the centre is $X \cup Y$—where X and Y are two classes—is included in the union of the two balls of which the centres are respectively X and Y:

$$B(X \cup Y, \rho) \subset B(X, \rho) \cup B(Y, \rho) \tag{8}$$

where $B(Z, \rho)$ is the set of classes of which the δ distance to Z is less than or equal to ρ.

The property (8) is satisfied for the most 'interesting' aggregation indices δ. It permits us appreciably to reduce the number of comparisons for detecting the class pairs which realize the minimal value of δ.

The 'reciprocal nearest neighbours' algorithm consists of joining—at each step—the class pairs $[X, Y]$ such that X (respectively Y) is the nearest neighbour of Y) (respectively X). One proves that this algorithm leads to exactly the same hierarchy of subsets of $\mathcal{O}$ as that obtained by the classical algorithm of hierarchical ascendant construction, provided that δ has the reducibility property.[3,12,13,14] The tree levels of the obtained tree can be retrieved from the hierarchy of subsets and from the values of δ concerned. The management of the transitivity of the 'reciprocal nearest neighbours' relation, is not trivial.[13]

The most recent work of M. Bruynooghe[3] consists of simultaneously exploiting the reducibility property and the 'reciprocal nearest neighbours' algorithm. Otherwise, the complexity simplification in space and time of a given algorithm of hierarchical classification may depend on the possible representation of the object set $\mathcal{O}$, by a cloud of points in a Euclidean space which provides the distance d.

One important idea consists of introducing parallelism into the hierarchical classification. The parallelism procedure is conceived at a global level.[13,15] According to a random sampling, we partition the set $\mathcal{O}$ to be classified into 'blocks' of the same size, to one unity exception. This size is the largest possible, taking into account the processing capacity of the computing system, on which a given hierarchical classification algorithm has to be used. We apply in parallel the latter to the different blocks. For each of them, we stop the evolution of the

algorithm as soon as we obtain a partition with a number of classes of a given order and statistically significant (using a statistical criterion). From this stage and by means of the same algorithm, the classification process is carried out on the set of the classes formed through the different blocks.

The final result is an approximation of the exact tree which would have been obtained by application of the algorithm on the whole set $\mathcal{O}$. This approximation has to be studied from a statistical point of view.

Problem 3: Analytical study of the algorithm behaviour (applied to a random sample) provided by a multidimensional distribution

Indeed—for given d and δ—we do not know how to set up in a relevant way this problem of statistical convergency. Nevertheless, several important studies have been performed in the case of the algorithm family of non-hierarchical classification that we call 'reallocating–recentring' (k-means and extensions).[16–20] We are going to end this section by some general remarks on the classical view of the context in which the ascendant hierarchical classification is set up.

In such a view, the representation of a given element of the set E to be classified is a point—eventually weighted—of a geometrical space. But such a representation fits clearly only in the case of the description of a set $\mathcal{O}$ of objects by a set V of quantitative or numerical variables.

The correspond analysis coding adapts to this view the mathematical representation of the set of rows of a contingency table.

It is quite common in France to classify the set $\mathcal{O}$ of objects on the basis of the description by the first components of a factorial analysis (e.g. *PCA*) of the cloud of points associated with $[\mathcal{O}, \mu_{\mathcal{O}}]$ in a Euclidean space. This practice is interesting in itself and can be entirely justified in the context of factorial analysis, which supposes the criterion of explained inertia. But this practice is outside the principles of taxonomy, where the different descriptive characters have to be considered *a priori* with the same importance. Apart from the fact of lack of intelligibility, this practice becomes mentally dangerous if it is systematically considered.

Otherwise, the classical approach forgets the following two fundamental points:

(*a*) The description by quantitative variables corresponds to one type of data. Even this type of data has directly to be considered, the significant descriptive information, contained in a given variable—for classificatory 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 preordonnance variable).

(*b*) The problem of the classification of the set V of descriptive variables is completely ignored. But the hierarchical classification of V into 'significant' classes and subclasses 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.

Let us come back to the mathematical simplicity of the algorithm of 'top-down hierarchical construction' of the classification tree.

Scientific advantage

The flexibility of use enables us to take into account multiple types of data descriptions.

Danger

The 'significance' of the automatic synthesis depends on two crucial aspects:

(i) the 'relevant coding' of information, results from experimental data or knowledge data;
(ii) the 'relevant notion of proximity' on the set $P(E)$ of all subsets of the set E to be classified.

3. THE 'LIKELIHOOD LINKAGE ANALYSIS (LLA)' CLASSIFICATION METHOD

3.1. General concepts and probabilistic association

The LLA hierarchical classification method concerns any mathematical or logical type of data table T crossing a set $\mathcal{O}$ of objects (respectively C of concepts) with a set a of descriptive 'variables' (we also say 'attributes' or 'parameters').

The method enables us to classify the set a of the variables[21-24] AND to classify the set $\mathcal{O}$ of the objects[21,25,13,26] (respectively C of the concepts[21,27-28]).

The data table T can be provided as well as by experimental observation or by knowledge. Indeed, given the relational system of Tarski:[29]

$$T = \langle \mathcal{O}; R_1, R_2, \ldots, R_j, \ldots, R_m \rangle \tag{9}$$

where the relations $R_j, 1 \leqslant j \leqslant m$, are supposed to have the same combinatoric type, the method is able to handle the two dual problems:

(i) classification by proximity of the relation set $\{R_j / 1 \leqslant j \leqslant m\}$ observed on $\mathcal{O}$
(ii) classification by proximity of the object set $\mathcal{O}$ described by the set $\{R_j / 1 \leqslant j \leqslant m\}$ of relations that we consider *a priori* to have the same importance.

As a matter of fact, we consider a descriptive attribute of an object set $\mathcal{O}$ as defining a q-ary relation on $\mathcal{O}$ that we represent by a structured subset of $\mathcal{O}^q$. In practice, it was sufficient to consider $q = 1$ or 2 or 4, in the qualitative or quantitative data analysis that we have dealt with. Nevertheless, the computing aspect concerning our association coefficients between relations on $\mathcal{O}$, have recently been resolved for very general cases. On the other hand, the relations $R_j, 1 \leqslant j \leqslant m$, can be weighted, in our treatment.

Let us now introduce another type of system that we denote by

$$S = \langle C; R_1, R_2, \ldots, R_j, \ldots, R_m \rangle \tag{10}$$

where C is a set of concepts or classes. Herein the data correspond to the statistical distribution of each R_j, $1 \leqslant j \leqslant m$, on each concept c belonging to C.

Also, for this data structure the LLA method assumes the two dual problems:

(i) classification by proximity of the set of the relations $\{R_j / 1 \leqslant j \leqslant m\}$ observed from their respective statistical distributions on the different concepts c of C;
(ii) classification by proximity of the concept set C described by the statistical distributions of each of the relations $R_1, R_2, \ldots, R_m$, considered *a priori* with the same importance.

We may note that the data structure concerned by correspondence analysis (contingency table or even horizontal juxtaposition of contingency tables) is a particular case of the system S.

The system S enables us to formalize and to treat a knowledge base elaborated by J. Lebbe, J.-P. Dedet and R. Vignes (1987). To this purpose we have introduced a new type of description variable called the 'taxonomic preordonnance variable with multiple choice'.[28] A 'preordonnance' variable is a qualitative variable of which the set of modality couples is provided by a total preorder which expresses—in an ordinal manner—the similarity perception between the variable modalities.[13,26] A 'taxonomic' variable[26] can be interpreted as a particular case of the 'preordonnance' variable. A 'taxonomic preordonnance variable with multiple choice'[28] is generally obtained from a hierarchical organization of preordonnance variables, where the 'value' of a given qualitative variable on a given concept is a logical formula on the modality set of the variable.

Example

Let us consider the variables 1, 18, 19 and 20 of the above mentioned data base knowledge concerning the biological descriptions of phlebotomine sandflies of French Guiana. a^1 is the sex, a^2 is the number of style spines, a^3 is the distribution of the insertion of 4 style spines and a^4 is the distribution of the insertion of 5 style spines. We obtain the taxonomic structure shown in Figure 2.

We are now going to give the general idea of the construction of the probabilistic proximity indices that we have set up. In this elaboration we have to distinguish the two systems T and S. We also have to consider the two dual problems:

(i) elaboration of a probabilistic association coefficient on $\{R_j / 1 \leqslant j \leqslant m\}$;
(ii) elaboration of a probabilistic similarity index on $\mathcal{O}$ (respectively C).

Let us consider (i). Given two relations R_j and R_k, $1 \leqslant j < k \leqslant m$, we begin by considering a 'rough' index, denoted by $s(R_j, R_k)$. This index takes into account in detail the set theoretic representation of R_j (respectively R_k) that we denote by rep (R_j) (respectively rep (R_k)). rep (R_j)

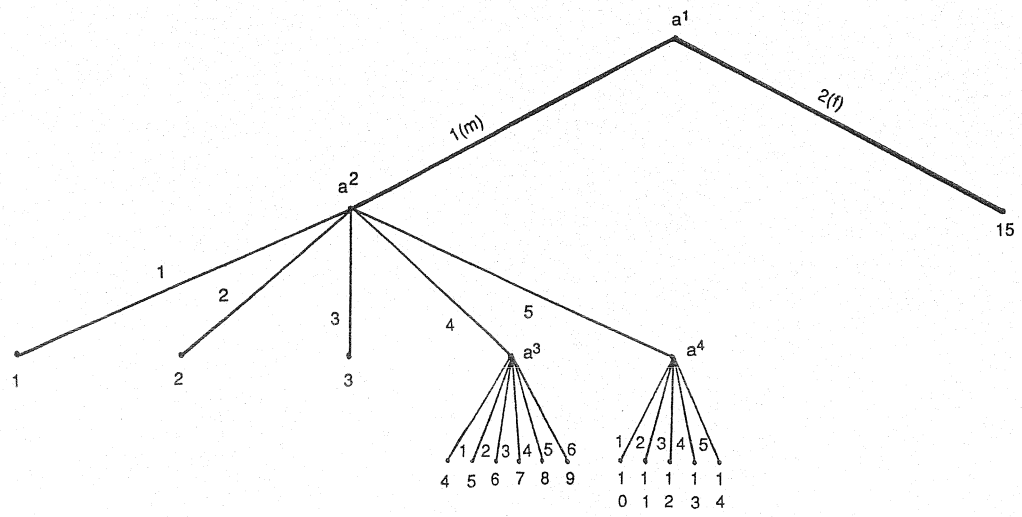

Figure 2.

(respectively rep (R_k)) is generally a structured part of $\mathcal{O}q$, or a weighting on $\mathcal{O}q$, if the relations are q-ary.

The following step consists of associating with the observed couple (R_j, R_k) of relations, a couple (R_j^*, R_k^*) of independent random relations, where R_j^*(respectively R_k^*) respects the combinatoric type of R_j (respectively R_k). Thus, for example, if R_j is a partition relation of the type (i.e. sequence of the class cardinals) t, R_j^*can be chosen as a random partition with fixed type t. This association is what we call 'hypothesis of no link (h.n.l.)' or more commonly 'hypothesis of independence'.

In the third step we mathematically compute the mean and the variance of the random 'rough' index $s(R_j^*, R_k^*)$, which we denote by $E[s(R_j^*, R_k^*)]$ and var $[s(R_j^*, R_k^*)]$. These calculations lead to what we call the 'locally' normalized coefficient

$$Q(R_j, R_k) = \frac{s(R_j, R_k) - E[s(R_j^*, R_k^*)]}{(\text{var } [s(R_j^*, R_k^*)]^{1/2}} \tag{11}$$

This coefficient permits us to reach the value of what we designates the 'local' probabilistic index; namely:

$$P(j, k) = \Pr \{s(R_j^*, R_k^*) \leqslant s(R_j, R_k)/\text{h.n.l.}\} \tag{12}$$

The reason why we are able to compute $P(j, k)$ is the normal approximation of the probability distribution of $s(R_j^*, R_k^*)$ under the h.n.l. This approximation is justified in most cases. Then,

$$P(j, k) \simeq \Phi [Q(R_j, R_k)] \tag{13}$$

where Φ denotes the cumulative normal distribution function $N(0, 1)$.

The formal analysis of $Q(R_j, R_k)$[30-33]—from an inferential point of view where the object set $\mathcal{O}$ is an increasing random sample with universe Ω—shows that $P(j, k)$ tends rapidly to zero (respectively unity) in the case of positive (respectively negative) association, between R_j and R_k on Ω. Thus, not all the richness of the probability scale can be exploited by means of the formula (13).

The point is that the coefficient $Q(R_j, R_k,)$ does not take into account the relative value of the association between R_j and R_k, with respect to the mutual association coefficient

$$\{Q(R_g, R_h)/1 \leqslant g < h \leqslant m\} \tag{14}$$

One way to make reference to (14), is to substitute to $Q(R_j, R_k)$, the standardized index

$$Q_s(R_j, R_k) = \frac{Q(R_j, R_k) - m_e(Q)}{(\text{var}(Q))^{1/2}} \tag{15}$$

where $m_e(Q)$ and $\text{var}_e(Q)$ are respectively the empirical mean and variance of (14).

The reference to a probability scale makes use of the limit distribution of $Q_s(R_j^*, R_k^*)$, under the independence hypothesis, which globally associates with the $\{R_j/1 \leqslant j \leqslant k\}$ of observed relations a family $\{R_j^*/1 \leqslant j \leqslant k\}$ of independent relations. In this correspondence $R_j^*$is of the same combinatoric type as R_j, respecting the cardinal characteristics of R_j.

Except in few structural situations, the limit distribution of $Q_s(R_j^*, R_k)$ is normal $N(0, 1)$. Then, the definitive probabilistic association coefficient can be put in the following form:

$$P_s(j, k) = \Pr \{Q_s(R_j^*, R_k^*) \leqslant Q_s(R_j, R_k)/\text{h.n.l.}\} \tag{16}$$

and computed—in most cases—by the following approximation:

$$P_s(j, k) \simeq \Phi [Q_s(R_j, R_k)] \tag{17}$$

$1 \leqslant j < k \leqslant m$, where Φ denotes the cumulative normal distribution function $N(0, 1)$, By this method, the probability scale is finely discriminant for the mutual organization of the set of the relations R_j, $1 \leqslant j \leqslant k$.

The problem of establishing a probability scale for the mutual comparison between the elements of the object set $\mathcal{O}$ (respectively the concept set C) is the dual of the preceding one. Let us consider the comparison between two objects x and y with respect to the comparison two by two within the object set $\mathcal{O}$. Our method in establishing a probabilistic similarity index on $\mathcal{O}$ can be decomposed as follows.[26]

(i) For each $j, 1 \leqslant j \leqslant m$, define a 'rough' similarity index $r_j(x, y)$, between x and y relative to the relation R_j.

(ii) Statistically normalize $r_j(x, y)$ with respect to the empirical distribution of r_j, on the object couples $\mathcal{O} \times \mathcal{O}$, to obtain

$$S_j(x, y) = \frac{r_j(x, y) - m_e(r_j)}{(\mathrm{var}_e\,(r_j))^{1/2}} \tag{18}$$

where $m_e(r_j)$ and $\mathrm{var}_e\,(r_j)$ are respectively the mean and the variance of r_j on $\mathcal{O} \times \mathcal{O}$. We call $S_j(x, y)$ the normalized contribution of R_j in the comparison of x and y.

(iii) Consider the sum over $\{j/1 \leqslant j \leqslant m\}$ of the normalized contributions to obtain:

$$S(x, y) = \sum_{1 \leqslant j \leqslant m} S_j(x, y) \tag{19}$$

(iv) Statistically normalize $S(x, y)$, with respect to the empirical distribution of S, on the object pairs $P_2(\mathcal{O})$, to obtain

$$Q_s(x, y) = \frac{S(x, y) - m_e(S)}{(\mathrm{var}_e\,(S))^{1/2}} \tag{20}$$

where $m_e(S)$ and $\mathrm{var}_e\,(S)$ are respectively the empirical mean and variance of S on $P_2(\mathcal{O})$.

(v) Compute the probability index

$$P_s(x, y) = \Pr\,\{Q_s^*(x, y) \leqslant Q_s(x, y)/\text{h.n.l.}\} \tag{21}$$

where $Q_s^*(x, y)$ is the random index associated with $Q_s(x, y)$ under the above h.n.l. which associates with the family of observed relations $\{R_j/1 \leqslant j \leqslant m\}$, a family of random relations $\{R_j^*/1 \leqslant j \leqslant m\}$. In order to compute $P_s(x, y)$, we use the normal approximation, namely

$$P_s(x, y) = \Phi[Q_s(x, y)] \tag{22}$$

where Φ is the cumulative normal distribution function $N(0, 1)$. This approximation can be justified by asymptotic sampling theory, in the case where m is large enough.

At this stage we may suppose that whatever the nature of the set E to be classified ($E = A$ or $E = \mathcal{O}$ (respectively C), according to the notation of the beginning of this section), we have a probabilistic similarity table.

$$\{P_s(x, y)/\{x, y\} \in P_2(E)\} \tag{23}$$

where $P_2(E)$ is the set of unordered pairs of elements of E. On the other hand, each element of the preceding table becomes, under the h.n.l., a random uniform distributed variable on the interval $[0, 1]$.

3.2. Family of criteria of the 'maximal link likelihood'

In order to understand in an intuitive manner this type of proximity criterion between classes, we consider Figure 3, which is placed in the simple and classical context of a two dimensional cloud. We suggest two class pairs; $\{C_1, C_2\}$ on 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

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

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

We have

$$\delta(C_1, C_2) < \delta(D_1, D_2) \tag{25}$$

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.

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 and 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 Section 3.2, the starting point of the construction

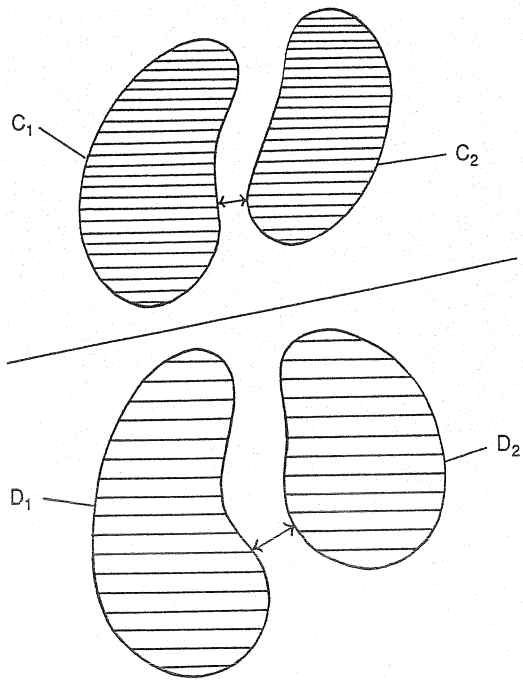

Figure 3.

of the aggregation criterion between two classes C and D (disjoint parts of E), which is called the 'maximal link likelihood', is:

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

to the couple (C, D), we associate a couple (C^*, D^*) of independent random classes, where C^* (respectively D^*) is composed of independent elements relating 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 (cf. section 3.2) 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

$$\begin{aligned} P(C, D) &= \Pr \{p(C^*, D^*) \leqslant p(C, D) / H\} \\ &= [p(C, D)]^{l \times m} \end{aligned} \tag{27}$$

where $l = \text{card } (C)$ and $m = \text{card } (D)$.[34]

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

$$VL_1(C, D) = [p(C, D)]^{l \times m} \tag{28}$$

Following the work of Nicolaü (1981)[35,36], we have suggested the family of criteria:

$$VL_\varepsilon(C, D) = [p(C, D)]^{(l \times m)^\varepsilon} \tag{29}$$

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:

$$S_\varepsilon(C, D) = -\log \{-\log [VL_\varepsilon(C, D)]\} \tag{30}$$

which leads to exactly the same classification tree.

We do have a reactualization formula for S_ε in case of multiple aggregation.[11] Otherwise, whatever the ε value ($0 \leqslant \varepsilon \leqslant 1$), the criterion S_ε possesses the reducibility property. In fact, the higher $\varepsilon (0 \leqslant \varepsilon \leqslant 1)$, the stronger is the reducibility property of S_ε.

3.3. Different stages of data analysis by the LLA method

Above LLA_ε

Stage 1. Collecting data;

Stage 2. Data coding: extraction of a data table $\mathcal{O} \times A$ or $C \times A$, and eventually, extraction of a system of weights (i.e. a measure) $\mu(\mathcal{O})$ on $\mathcal{O}$, or $\mu(C)$ on C (see for the notation, the first passage of Section 3.1).

Stage 3. Mathematical representation (set theoretic and combinatorics representation, even in case of quantitative variables).

In fact, we have seen above that we interpret a descriptive attribute in terms of a relation which is—in the cases that we have studied as yet—unary or binary or quaternary. The quantitative case is represented by a weighted relation. Thus, if the description concerns a set

$\mathcal{O}$ of elementary objects, the attribute representation is a structured subset—weighted eventually—of $\mathcal{O}$ or $\mathcal{O}\times\mathcal{O}$ or $(\mathcal{O}\times\mathcal{O})\times(\mathcal{O}\times\mathcal{O})$.

Stage 4. Table of proximity indices P_s between the elements of the set E to be classified (cf. Reference 18).

Below LLA_ε

Stage 5. Detection of the 'significant nodes' of the classification tree. This detection is based on the elaboration of an association criterion which matches a given partition π and an ordinal information $\omega(E)$,[25,37] relative to the resemblances between the elements of the set E to be classified. For this purpose, we interpret a partition π as inducing a total pre-order into two classes on the set $F=P_2(E)$ of unordered object pairs: $S(\pi)$ and $R(\pi)$, where $S(\pi)$ (respectively $R(\pi)$) is the set of separated (respectively joined) pairs. We set

$$S(\pi) < R(\pi) \tag{31}$$

On the other hand, $\omega(E)$ is a total pre-order on F, associated to the similarity index P_s:

$$[\forall(p,q)\in F\times F],\quad p\leqslant q \Leftrightarrow P_s(p)\leqslant P_s(q) \tag{32}$$

In this way, one finds oneself faced with the comparison of two combinatoric structures of the same nature (total pre-orders on F). For this comparison—as expressed previously—we have built an association coefficient which is statistically normalized (compare with expression (11), above) and which we denote here by $S[\omega(E),\pi]$. Let us consider now the two following sequences:

$$\{S[\omega(E),\pi_i]/1\leqslant i\leqslant I\} \tag{33}$$

and

$$\{\tau_i = S[\omega(E),\pi_i]-S[\omega(E),\pi_{i-1}]/2\leqslant i\leqslant I\} \tag{34}$$

where π_i is the emerged partition at the ith level of the classification tree and where I is the total number of levels.

A 'significant' node corresponds to a local maxima of the distribution (34) of the increasing rate of the criterion $S[\omega(E),\pi)$, on the increasing sequence of tree levels.

Stage 6. Reduction of the classification tree to a system N of nodes, such that each n of N is either significant, or a father of a significant node.

4. GENERAL SCHEME OF TREATMENT

—DO: $E=A$
$ARBA = LLA(E)$

—DO: $E=O$ (respectively C), according to the data table nature.
$ARBO = LLA(E)$

If $E=A \Rightarrow$ set : dual $(E)=\mathcal{O}$ (respectively C), according to the data table nature.

If $E=\mathcal{O}$ (respectively C) $\Rightarrow$ 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), with respect to an organized system of classes on a subset of dual (E) (respectively E).[38–42]

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.

REFERENCES

1. M. Anderberg, *Cluster Analysis for Applications*, Academic Press, New York, 1973.
2. J. P. Benzecri, *L'analyse des Données'*. Tome 1: *La Taxinomie*, Dunod, Paris, 1973.
3. M. Bruynooghe, 'Nouveaux algorithmes en classification automatique applicables aux très grands ensembles de données, recontrées en traitement d'images et en reconnaissance des formes', Thèse de Doctorat d'Etat, Université de Paris VI. 23, janvier 1989.
4. E. Diday, J. Lemaire, J. Pouget and F. Testu, *'Eléments d'Analyse de Données'* Dunod, Paris, 1982.
5. J. A. Hartigan, *Clustering Algorithms*, Wiley, New York, 1975.
6. M. Jambu, *Exploration Informatique et Statistique des Données'*, Dunod, Paris 1989.
7. N. Jardine and R. Sibson, *Mathematical Taxonomy*, Wiley, London, 1971.
8. P. H. A. Sneath and R. R. Sokal, *Numerical Taxonomy*, W. H. Freeman and Company, San Francisco, 1971.'
9. G. N. Lance and W. T. Williams, 'A general theory of classification sorting strategies: 1 = hierarchical systems, 2 clustering systems', *Computer Journal*, 9–10, No. 4, 373–380, No. 3, 271–276 (1967).
10. M. L. Tricot and M. Donegani, 'Présentation of unifiée des indices de proximité entre classes en classification hiérarchique ascendante' *R.A.I.R.O.*, **23** (2), pp. 165–192, 1989.
11. I. C. Lerman 'Formules de réactualisation en cas d'agrégations multiples', *R.A.I.R.O.*, **23**(2), 151–163 (1989).
12. J. Juan, 'Le programme HIVOR de classification ascendante hiérarchique selon les voisins réciproques et le critère de la variance', *Cahier de l'Analyse des Données*, vol. 7, 1982 pp. 173–184.
13. Ph. Peter, 'Méthodes de classification hiérarchique et problèmes de structuration et de recherche d'informations assistées par ordinateur', Thèse de l'Université de Rennes I, 6 mars 1987.
14. C. de Rham. 'La classification hiérarchique ascendante selon la méthode des voisins réciproques', *Cahiers de l'Analyse des Données*, **5**(2), 135–144 (1980).
15. I. C. Lerman and Ph. Peter, 'Organisation et consultation d'une banque de petites annonces à partir d'une méthode de classification hiérarchique en paralléle', in 'Data Analysis and Informatics 4' North Holland (Diday 8 al. eds), 121–136, 1986.
16. G. Celeux, 'Reconnaissance de mélanges de densités de probabilité et applications à la validation des résultats en classification', Thése de Doctorat d'Etat, Université de Paris IX Dauphine, septembre 1987.
17. J. Diebolt, J. C. Simon and W. L. Miranker, 'The dynamic cluster algorithm with continuous data', Research Report RC 6743, I.B.M. Research Division, 1977.
18. J. Lemaire, 'Etude de propriété asymptotiques en classification', Thèse de Doctorat d'Etat, Université de Nice, janvier 1990.
19. I. C. Lerman, 'Convergence optimale de l'algorithme de "réallocation-recentrage" dans le cas continu le plus simple', *R.A.I.R.O.*, **20**(1) 19–50 1986.
20. D. Pollard, 'Strong consistency of k-means clustering', *Annals of Statistics*, **9**, 135–140 (1981).
21. I. C. Lerman, *Classification et Analyse Ordinale des Données*, Dunod, Paris, 1981.
22. I. C. Lerman, 'Indices d'association partielle entre variables qualitatives nominales', *R.A.I.R.O. Série Verte*, **17**(3), 213–259 (1983).
23. I. C. Lerman, 'Indices d'association partielle entre variables qualitatives ordinales', *Publications d'Institute de Statistiques de l'Université de Paris XXVIII*, fasc. 1(2), 7–46 (1983).
24. I. C. Lerman, 'Comparing relational variables according to likelihood of the links classification method', Japanase–French Scientific Seminar, Tokyo, March 24–26 1987, In E. Diday, C. Hayashi, M. Jambu and N. Ohsumi (eds), *Recent Developments in Clustering and Data Analysis*, Academic Press, New York, 1988.
25. I. C. Lerman, *Les Bases de la Classification Automatique*, Gauthier Villars, Paris, 1970.
26. I. C. Lerman, 'Construction d'un indice de similarité entre objets décrits par des variables d'un type quelconque. Application au problème du consensus en classification', *Revue de Statistique Appliquée*, XXXV(2), 39–60 (1987).

27. B. Tallur, 'Contribution à l'analyse exploratoire de tableaux de contingence par la classification', Thèse de Doctorat ès Sciences, Univ. de Rennes I, 23 septembre 1988.
28. I. C. Lerman and Ph. Peter, 'Classification of concepts described by taxonomic preordonnance variables with multiples choice. Application to the structuration of a species set of phlebotomine', in E. Diday, INRIA (ed), '*Data analysis, Learning symbolic and Numerical knowledge*', Nova Science Publishers, 1990, pp. 73–85.
29. A. Tarski, 'Contribution to the theory of models', *I,II. Indagationes Mathematicae*, **16**, 572–588 (1954).
30. I. C. Lerman, 'Justification et validité statistique d'une échelle [0, 1] de fréquence mathématique pour une structure de proximité sur un ensemble de variables observées, IRISA, PI no. 221, janvier 1984, 47, pp. paru dans *Publications de l'Institute de Statistique de l'Université de Paris, XXIX*, fasc. 3(4), 27–57 (1984).
31. I. C. Lerman, 'Analyse de la forme limite de coefficients statistiques d'association entre variables relationnelles', Rapport IRISA no. 367, juin 1987.
32. F. Daudé, 'Normalisation sous hypothèse d'absence de lien', Publication interne IRISA no. 549, septembre 1990, 42 pp.
33. M. Ouali Allah, 'Analyse de la forme d'un coefficient d'association entre variables qualitatives', Publication interne IRISA no. 554, octobre 1990, 26 pp.
34. I. C. Lerman, 'Sur l'analyse des données préalablee à une classification automatique. Proposition d'une nouvelle mesure de similarité', *Revue Mathematique et Sciences Humaines*, **32**, (1970).
35. F. Nicolaü, 'Criterios de analise classificatoria hierarquica baseados na funçao de distribuiçao', Thése de Ph.D. Faculté des Sciences de Lisbonne, 24 février 1981.
36. F. Costa Nicolau and M. P. Brito, 'Improvements in Nhmean method', in E. Diday, INRIA (ed.) *Data Analysis. Learning Symbolic and Numerical Knowledge* Nova Science Publishers, New York, 1989, pp. 73–87.
37. I. C. Lerman, 'Sur la signification des classes issues d'une classification automatique', in J. Felsenstein (ed.), *Numérical Taxonomy*, NATO ASI Series Vol. G1, Springer-Verlag, Berlin, 1983, pp. 179–198.
38. J. P. Geffrault, 'Discrimination de classes et détermination d'ensembles minimaux de mesures par la classification automatique de formes', Thèse de Doctorat de 3ème cycle, Université de Rennes I, 15 mars 1982.
39. I. C. Lerman, 'Association entre variables qualitatives ordinales nettes ou floues', *Revue Statistique et Analyse des données*, **8** (7) 41–73 (1983).
40. I. C. Lerman, M. Hardouin and Th. Chantrel, 'Analyse de la situation relative entre deux classifications floues, Secondes journées internationales Analyse des Données et Informatique, Versailles 17–19 octobre 1979, paru dans *Data Analysis and Informatics*, North-Holland 1980, pp. 523–533.
41. A. Moreau, 'Elaboration et calcul d'indices d'association entre variables qualitatives "nettes" ou "floues". Application à une forme d'interprétation d'une classification de paramètres épidémiologiques', Thèse de Doctorat de 3ème Cycle, Université de Rennes I, 14 Juin 1985.
42. A. Prod'homme, 'Indices d'explication des classes obtenues par une méthode de classification hiérarchieque respectant la contrainte de conttiguïté spatiale. Application à la viticulture Girondine et à la construction de logements dans les Bouches-du-Rhône', Thèse de Doctorat du 3ème Cycle, Université de Rennes I, 19 décembre 1980.