DESCRIPTION CONTRASTING
in
INCREMENTAL CONCEPT FORMATION

Christine DECAESTECKER*

CADEPS Artificial Intelligence Research Unit, Université Libre de Bruxelles
Av. F.D.Roosevelt 50, C.P. 194/7, B-1050, Brussels, Belgium
Tel (+ 32 2) 650 2783; Email CADEPS@BBRNSF11(.BITNET)

ABSTRACT

This study evaluates the impact of concept descriptions on the behaviour and performance of concept formation processes (in which the data is either noisy or noise-free). Using a common architecture (ADECLU), different concept definitions are envisaged. These descriptions are of symbolic/numeric type, including statistical indices. The use of these indices introduces a "contrasting" between concept descriptions and reduces the effect of noise on predictive performance.

KEYWORDS : Concept Formation, Incremental Conceptual Clustering, Attribute selection, Typical value.

1. INTRODUCTION

The general aim of incremental concept formation (Gennari & al, 1989) is to construct (on the basis of sequentially presented object descriptions) a (hierarchical) classification of objects where each class is provided with a definition which summarises its elements. Further aims are to condense (to summarise) information, to use the knowledge thus obtained to classify new objects, and to make predictions concerning unknown values of the objects. Thus, the performance of the acquired hierarchies is (predominantly) measured in terms of their ability to make predictions concerning unknown attributes. This ability is usually called the *"predictive power"* of the hierarchies (Gennari & al, 1989).

* This work is supported by the Belgian National incentive-program for fundamental research in Artificial Intelligence. The scientific responsibility is assumed by the author.

In earlier work (Decaestecker, 1989a,b), the author studied Incremental Concept Formation and developed ADECLU, an incremental hierarchical Conceptual Clustering system, which extends ideas from UNIMEM (Lebowitz, 1987) and COBWEB (Fisher, 1987). The aim of our research was to propose a process which selects attributes which best describe a concept.

A simple version (ADECLU1) (Decaestecker, 1989a) keeps only those characteristics which are common to the objects of the same class. A second version (ADECLU2)(Decaestecker, 1989b) keeps for each class, all the values observed, but uses association criterion from Data Analysis (especially the Goodman-Kruskal "Lambda" index), to select combinations of attributes which are more appropriate to describe a concept.

A new version (ADECLU/S) improves concept description by selecting, for each attribute, a set of typical values from all the values observed in the concept. The general aim is to obtain "contrasted" concept descriptions. This selection of typical values is also based on the optimisation of a statistical index (which guarantees good statistical properties to the selected values), and occurs together with the selection of the characteristics already used in ADECLU2. We observe that ADECLU/S seems better than ADECLU2, in recognizing and differentiating concepts with inter-class proximity (i.e. the elements of one class are close to the elements of another class), for which description contrasting is then useful. In this paper, we wish to show the influence of the knowledge representation (i.e. the concept descriptions) on the predictive power of the hierarchies produced by incremental processes such as ADECLU, in the presence and absence of noise in the data. Another point already mentioned in (Decaestecker, 1989a,b & 1990), concerns the quality of classification of the data (from the point of view of traditional Data Analysis).

Section 2 proposes a general description of ADECLU. In section 3, the formalism is specified for each of the 3 versions. This section aims to clarify the evolution of the knowledge representation between the different versions, as well as their inherent statistical properties. Section 4 clarifies a "cutoff" technique of the hierarchy. Section 5 presents the results obtained in the absence and present of noise for the different versions.

2. GENERAL DESCRIPTION

In this section, we recall the different characteristics of ADECLU, in a general formalism which is suitable for the 3 versions.

2.0 Description and Organisation of the Objects

The objects are described by a conjunction of attribute/value pairs. We will only treat qualitative attributes (or data translated into this qualitative format).

ADECLU organises a series of objects presented sequentially, into a hierarchy in which each node represents a concept and its description (only those descriptions occuring in the classification operation). The set of objects classified under a concept *C* is partitioned into the subconcepts of *C*.

2.1 Description of the Concepts

Each concept is described by a conjunction of characteristics, where :

characteristic = quadruplet : $(j\ ,\ V_j\ ,\ {}^cV_j\ ,\ w_j)$

where j is the attribute.

V_j is the set of values (of the attribute j) retained to describe (the objects integrated into) the concept.

${}^cV_j$ is the set of values (of the attribute j) observed in the concept but not retained to describe it.

w_j is the weight ($0 \leq w_j \leq 1$) which is a function of the information added by the characteristic to the concept definition. This weighting, called *"relevance"* of the characteristic, is expected to vary during the incremental process.

Later, we shall specify this general formalism for each of the versions of ADECLU. We shall see that for the first two versions (ADECLU1 and ADECLU2), the selection of V_j is arbitrary (but consistent with common sense). However, the third version (ADECLU/S) performs a "statistical" selection of the values (called "typical") for the concept considered.

ADECLU is also characterised by its selection of conjunctions of relevant characteristics to describe a concept. This selection of attributes is realized by the weights w_j which distinguish the *relevant* characteristics ($w_j > 0$) for classification and prediction, from the characteristics which are merely *descriptive* ($w_j = 0$). This selection is arbitrary in ADECLU1 and "statistical" in ADECLU2 and ADECLU/S. Hence the concepts provided by these two version can be labelled *"statistical"*. They offer an interesting alternative to "probabilistic" concepts. The probabilistic concepts produced by COBWEB incorporate in the concept description, the conditional probabilities $P(O \in C \mid j(O) = v)$ and $P(j(O) = v \mid O \in C)$ ("predictiveness" and "predictability" of a value v for a concept C) for each value of each attribute. As we shall see (section 3.2), the weights w_j condenses these two notions and generalises them to the set V_j of selected values for an attribute j in a concept C. One advantage of this quality measure is the possibility to eliminate the inadequate internal disjunctions in a concept description ($w_j = 0$). An other is to can construct an internal disjunction of possible (but not all) values which optimises this measure. This possibility to increase the relevance of the characteristics, has motivated the selection of typical values in ADECLU/S.

2.2 Suitability criterion (evaluation function)

As with most Concept Formation systems (Fisher, 1987; Gennari & al, 1989), ADECLU uses a hill climbing search strategy in a space of concept hierarchies. When a new object O is presented, a search is undertaken to find the most specific concept to classify the new object, based on the descriptions of the different concepts (partial matching). At each level of the hierarchy, a series of operators (creation, split, merge) is applied to an initial partition. The resulting partitions are compared using an evaluation function. The best concept and the corresponding partition are selected. Then, the process recurses for the subconcepts of the concept selected at the previous step.

We now present the evaluation function (used in ADECLU) called the *Suitability Criterion.* It is a context-dependent measure which is a function of the object description, the concept description and the partition which contains the concept.

Definition : The suitability S between an object O and a concept C_l of a partition P, is defined as follows:

$$S(O,C_l,P) = sc(O,C_l,P) - \Sigma_{k\neq l}\, sc(O,C_k,P)$$

where the score sc between an object O and a concept C_l is defined as

$$sc(O,C_l,P) = f(|C_l|)\, \Sigma_j\, p^O_{jl}\, w_{jl}$$

with $f(|C_l|)$: a function of the cardinal of C_l (i.e. the number of objects in C_l)

w_{jl} : the weighting of the characteristic $(j, V_{jl}, {}^cV_{jl}, w_{jl})$

p^O_{jl} : the matching factor which takes the value +1 or 0 or -1, depending on whether the attribute value of O (i.e. $j(O)$) and the attribute value of C_l ($V_{jl} \cup {}^cV_{jl}$), agree ($j(O) \in V_{jl}$) or are indifferent ($j(O) \in {}^cV_{jl}$) or disagree ($j(O) \notin (V_{jl} \cup {}^cV_{jl})$).

The factor $f(|C_l|)$ can take different forms. Details concerning its use and its effect on ADECLU2 can be found in (Decaestecker, 1989b).

3. EVOLUTION of the KNOWLEDGE REPRESENTATION

3.1 Specialising the general formalism for the different versions

The TABLE 1 below shows the general formalism introduced earlier, for each of the 3 versions of ADECLU.

	V_j	${}^cV_j$	w_j
ADECLU1	$\bigcap_{O \in C} j(O)$	$\{\bigcup_{O \in C} j(O)\} \setminus V_j$	0 if $V_j = \emptyset$ 1 otherwise
ADECLU2	$\bigcup_{O \in C} j(O)$	$\emptyset$	$L(V_j,C)$
ADECLU/S	typical set	$\{\bigcup_{O \in C} j(O)\} \setminus V_j$	$L(V_j,C)$

TABLE 1 : 3 versions of ADECLU

ADECLU1 is the simplest of the versions, which initially can be used to show and to test the incremental strategy and the adequacy criterion presented in section 2.2. This simplified form produces good results on noise-free data (see section 5 and (Decaestecker, 1989a)), which justifies the general principles controlling the use of ADECLU. In ADECLU1, each concept C is simply defined by a conjunction of characteristics which are common to all its objects (with inheritance rule).

In ADECLU2, the description of concepts becomes more general and allows multivalued characteristics, defined simply by the values observed in the concept, for a given attribute. The weights w_j, in the definition of the concept C will reflect the degrees of association (in the Data Analytic sense (Andeberg, 1973)) which exist between the attribute j (more precisely the subset of observed values) and the fact of belonging to the concept C considered (cf section 3.2).

In the new version ADECLU/S, we introduce the notion of "typical values" from the values observed in the concept. The aim is to select, in each characteristic of a concept, a subset of typical values (V_j) which maximally "contrasts" the descriptions of "sybling" concepts (i.e. having the same parent in the hierarchy produced by ADECLU). The typicality of a value for a concept implies the notion of "distinctiveness" for this concept, i.e. that it will be observed in the majority of cases (with a superior frequency) in the concept relative to the others. This selection of values aims to increase the relevance of the characteristic and is based on a simple statistical index (cf section 3.3). With respect to the general formalism introduced in 2.1, only ADECLU/S actually uses ${}^cV_j$.

Both ADECLU2 and ADECLU/S use the value of an association criterion such as the definition of w_j. This criterion is the Lambda index of Goodman et Kruskal (represented as $L(V_j,C)$) between the memberships of V_j and C (cf section 3.2). This index seemed adequate to "summarise" two important aspects in the concept formation process, i.e. "classification" and "prediction of unknown values" (cf (Gennari & al, 1989)). More precisely, we are interested in:

1. the prediction that an object belongs to a concept C, given attribute values (for the classification of a new object)
2. the prediction that attributes have certain values, given that the object belongs to a concept C (for the prediction of unknown values)

We will see in the following section how the Lambda index (a symmetric optimal class prediction index) meets these requirements.We will study it in the more general case of ADECLU/S, where not all the observed values are kept to describe a concept. (The definitions and properties presented in this paper, generalize those of ADECLU2 presented in (Decaestecker, 1989b)).

3.2 Selection of relevant characteristics (ADECLU2 and ADECLU/S)

Goodman and Kruskal (1954) suggested measuring the association between two variables X and Y by the *predictive power* of one variable as a predictor of the other and is defined as the relative decrease in the probability of error (in the prediction of Y), due to knowledge of the value of X. If the prediction of the value of X from Y is as important in the model as Y from X, then these authors propose a symmetric relationship, by considering the prediction of X half the time, and Y the other half.

The probabilistic model is described in detail in (Goodman & al, 1954). The essence is as follows: For an object chosen at random in the population, one must predict at best, either the value of X or the value of Y (at random, with equal probabilities), given (case 1) no information or (case 2) the value of the other variable for this object (the value of X to predict Y and vice versa).

Definition : (Goodman & al, 1954)

$$\lambda = \frac{\text{(Prob. of error in case 1)} - \text{(Prob. of error in case 2)}}{\text{(Prob. of error in case 1)}}$$

If one only has a sample of the population (which is the case with incremental processes), one can estimate the value of λ from the contingency table between X and Y (table of absolute frequencies) from the sample :

matrix $(n_{ij})_{pxq}$ where n_{ij} is the number of objects for which $X = x_i$ and $Y=y_j$
with the marginal distributions $(n_{i.})$ and $(n_{.j})$ where $n_{i.} = \Sigma_j\, n_{ij}$ and $n_{.j} = \Sigma_i\, n_{ij}$

The maximum likelihood estimator of λ between X and Y is :

$$L = \frac{\Sigma_i\,(\max_j n_{ij}) + \Sigma_j\,(\max_i n_{ij}) - (\max_j n_{.j}) - (\max_i n_{i.})}{2\,n_{..} - (\max_j n_{.j}) - (\max_i n_{i.})}$$

In ADECLU2 and /S, we define the *relevance* (w_j) of the characteristics (when describing a concept) by the measure of the L indices between each attribute j and the membership of each concept C. Since we are dealing with incremental processes, we estimate, at each instant, the indices λ for the objects already processed (integrated into the hierarchy). Since these processes are subject to numerous updates, we estimate λ with a simplified 2x2 contingency table (see TABLE 2), between the variable "belonging to the concept" ($O \in C$?) and the variable "belonging to the set of values V_j in C for the attribute j ($j(O) \in V_j$?). In a hierarchical organisation, the set of objects which will be used as a reference to measure the association criterion (and on which will be constructed the contingency table), will consist only of those objects integrated into the parent of the concept C being considered.

In conclusion, for each characteristic of concept C, we will study the following contingency table :

$j(O) \in V_j$: $O \in C$:	yes	no	TOTAL
yes	$\lvert C\rvert - v$	v	$\lvert C\rvert$
no	$N1 - (\lvert C\rvert - v)$	$N - N1 - v$	$N - \lvert C\rvert$
TOTAL	$N1$	$N - N1$	N

where $|C|$ is the cardinal (card) of C (i.e. the number of objects in C), N is the card of the parent concept of C, $N1$ is the number of objects of the parent concept which match $j = V_j$ and v is the number of objects of C with $j(O) \notin V_j$.

TABLE 2 : Contingency table

The weight w_j is then the estimation L of the Lambda index applied to TABLE 2 :

$$w_j \equiv L = \frac{\max(|C|-v,\, v) + \max(N1-|C|+v,\, N-N1-v) + \max(|C|-v,\, N1-|C|+v) + \max(v,\, N-N1-v) - K}{2N-K}$$

where $K = \max(N1, N-N1) + \max(|C|, N-|C|)$

w_j therefore measures the increase in good prediction of "$\in C$" or "$\in V_j$", given the knowledge of the other variable, relative to the absence of any information.

Property 3.1 below guides the choice of L, among the possible criteria.

Property 3.1 : (2x2 table)

We have L=1 in the case of *"perfect association"* and *"perfect anti-association"* where the 2x2 tables have respectively the following forms :

$$\begin{array}{|c|c|}\hline a & 0 \\ \hline 0 & d \\ \hline \end{array} \qquad\qquad \begin{array}{|c|c|}\hline 0 & b \\ \hline c & 0 \\ \hline \end{array}$$

association — anti-association

We have L>0 in the cases of "good" association or "good" anti-association where :

$$\begin{array}{ccc} a & > & b \\ \vee & & \wedge \\ c & < & d \end{array} \qquad\qquad \begin{array}{ccc} a & < & b \\ \wedge & & \vee \\ c & > & d \end{array}$$

$(a > b\,, a > c\,, d > c\,, d > b)$ — $(b > a\,, b > d\,, c > a\,, c > d)$

association — anti-association

Also, $L > 0$ when one (and only one) of these relations is false.

We have L=0 when two of these relations are false. There are 4 possibilities (having the same relations in the two rows and in the two columns where, at most, two equalities are satisfied) which generalize the cases of independence (i.e. those having the same multiplicative factor between the two rows and between the two columns):

$$\begin{array}{ccc} a & \leq & b \\ |\!\wedge & & |\!\wedge \\ c & \leq & d \end{array} \qquad \begin{array}{ccc} a & \geq & b \\ |\!\wedge & & |\!\wedge \\ c & \geq & d \end{array} \qquad \begin{array}{ccc} a & \geq & b \\ |\!\vee & & |\!\vee \\ c & \geq & d \end{array} \qquad \begin{array}{ccc} a & \leq & b \\ |\!\vee & & |\!\vee \\ c & \leq & d \end{array}$$

The proof can be obtained by application of the definition of L.

Consequences for w_j :

For each characteristic $(j, V_j, {}^cV_j, w_j)$ of a concept C,

$w_j = 1$ when $j = V_j$ $(j \neq V_j)$ is a discriminating characteristic of C.

$w_j > 0$ explains (in the cases of association) a general tendency of the C elements to have $j(O)$ in V_j in contrast to the general tendency of the other elements (not in C) to have the value $j(O)$ out of V_j.

$w_j = 0$ when , for example,

a) $P(j(O) \in V_j \mid O \in C) \leq 1/2$ and $P(O \in C \mid j(O) \in V_j) \leq 1/2$
b) $P(j(O) \notin V_j \mid O \notin C) \leq 1/2$ and $P(O \in C \mid j(O) \in V_j) \leq 1/2$
c) $P(j(O) \notin V_j \mid O \notin C) \leq 1/2$ and $P(O \notin C \mid j(O) \notin V_j) \leq 1/2$
d) $P(j(O) \in V_j \mid O \in C) \leq 1/2$ and $P(O \notin C \mid j(O) \notin V_j) \leq 1/2$

When $w_j = 0$, it is fully justifiable to eliminate the characteristic relative to the attribute j from the definition of the concept C, because this characteristic is not relevant to the object classification nor to the attribute value prediction. Thus a characteristic of weight zero is not taken into account in the calculation of the score measuring the degree of "matching" between a new object and the concept being considered.

We must now distinguish (for $w > 0$) the cases of association from the cases of anti-association (between the two variables "$\in C$" and "$\in V_j$"). Since in ADECLU, the set V_j is formed from the observed values of attribute j in concept C, we will construct a set V_j *"in association"* with the concept C. But, in ADECLU2, there are no "anti-association" cases, because the value set V_j collects all the observed values and therefore the quantity "v" (of TABLE 2) is zero. In ADECLU/S, the selection of a set V_j in association with C is assured by the selection strategy itself.

3.3 Selection strategy of typical values in ADECLU/S

As announced in section 2.1, the selection values aims to increase the relevance of the characteristics. In TABLE 2 which defines the relevance w_j of a characteristic, this motivation correspond to construct a set V_j for which the quantities v and N_1–($\lvert C\rvert$–v) are as smaller as possible. This notion can be assimilated to the notion of "contrast" between concept descriptions.

The contrast between a concept C and its sibling concepts, with respect to the attribute j, is estimated by an index which is also defined using TABLE 2. This index (defined below) is called the *"jContrast"* (or simply jC) and is related to the Lambda index w_j. The jContrast is also the numerator of a measure credited to Hamann (Andeberg, 1973).

Definition :

$j(O) \in V_j$: $O \in C$:	yes	no	TOTAL
yes	a	b	$\lvert C\rvert$
no	c	d	$N-\lvert C\rvert$
TOTAL	$N1$	$N-N1$	N

where $a = \lvert C\rvert - v$
$b = v$
$c = N1 - \lvert C\rvert + v$
$d = N - N1 - v$

jContrast $= 1/2\,[(a-b) + (a-c) + (d-b) + (d-c)] = a + d - b - c$
$= N + 2\,\lvert C\rvert - 2\,N1 - 4\,v$

The aim of value selection is to construct (for each attribute j) a set V_j for which jContrast (jC) is maximum. One can see (by a simple but long calculation) that the optimization of jC corresponds generally with an increase of w_j. We have favoured the optimization of jC because it can be made easily and economically in the incremental process and gives, for each set V_j of selected values, good statistical properties.

We present now two properties of the jContrast index. The first shows that the sign of jC characterises the type of association in a 2x2 table. The second guides the optimisation of jC and characterises (see corollary) the set V_j finally selected by ADECLU/S. For more details and proofs, see (Decaestecker, 1990).

Property 3.2 :

For the 2x2 table before, we have:

$jC = 0$	when	$a + d = b + c$	in the case of "indetermination"
$jC > 0$	when	$a + d > b + c$	in the case of "association"
jC max $(= N)$	when	$b + c = 0$	in the case of "perfect association"
$jC < 0$	when	$a + d < b + c$	in the case of "anti-association"
jC min $(= -N)$	when	$a + d = 0$	in the case of "perfect anti-association"

Property 3.3 :

Let P be a set of disjoint concepts (a partition), and C a concept of P.
For each characteristic $(j, V_j, {}^cV_j, w_j)$ of C, if $jC \geq 0$,

the transfer of a value k of ${}^cV_j$ to V_j, increases the jC value if
"frequency of k in C" > "frequency of k in $P \backslash \{C\}$"

the transfer of a value k of V_j to ${}^cV_j$, increases the jC value if
"frequency of k in C" < "frequency of k in $P \backslash \{C\}$"

Definition : For each attribute j found in the description of a concept C, we assert that :
The selected set of values V_j or the division of the observed values by the pair $(V_j, {}^cV_j)$ is *optimal*, if the corresponding jC is maximum for this selection of values in the total set of observed values (of j in C).

Corollary of Property 3.3 :

If $(V_j, {}^cV_j)$ is *optimal* then each selected value k ($\in V_j$) verifies :

"frequency of k in C" > "frequency of k in $P \backslash \{C\}$"

which implies that (with K the number of concepts in partition P) :

"frequency of k in C" > $K/2$ "average frequency of k in a concept of P"

and more particularly

"frequency of k in C" = MAX ("frequency of k in C' ") (with $C' \in P$)

Let O be the new object that we have to classify, and let (C, P) be the pair for which (at the level considered) the suitability $S(O, C, P)$ is maximum. Thus ADECLU/S integrates the object O into concept C

(of the partition P). ADECLU/S must then adapt the description of C i.e. for each characteristic, the value set V_j and the weight w_j must be adjusted.

For each characteristic $(j, V_j, {}^cV_j, w_j)$ and the corresponding value $j(O)$ of O,
if $j(O) \in V_j$: first, we can leave V_j as it is,
and then, transfer OR not $j(O)$ from V_j to ${}^cV_j$
if $j(O) \notin V_j$: first, we can leave V_j as it is and insert $j(O)$ into ${}^cV_j$ (if $j(O) \notin {}^cV_j$),
and then, transfer OR not $j(O)$ from ${}^cV_j$ to V_j

We chose the case with jContrast maximum (and positive). It suffices to adjust the frequency of $j(O)$ in C and to effect the transfer if the corresponding inequality of property 3.3 is satisfied. Then the quantities $|C|$, $N1$, v are also updated andw_j can be calculated for the resulting typical set. In each sybling concept C' of $P\backslash\{C\}$, if $j(O) \in V_{jC'}$ and if the second inequality of property 3.3. is satisfied, then $j(O)$ must be transferred to ${}^cV_{jC'}$ (and the corresponding quantities $N1$ and v adjusted). In (Decaestecker, 1990), can be found more details of this strategy. It occurs by simple transfer between V_j and ${}^cV_j$ (for the concept considered and the sybling concepts) where the property 3.3 guides the choice.

4. CUTOFF IN THE CLASSIFICATION

In real world applications, a system which stores every object builds too large a hierarchy (to each object, corresponds a leaf of the hierarchy). Exhaustive trees, can have negative properties in noisy domains (Quinlan, 1986). To limit the number of concepts, several authors (e.g. (Gennari & al, 1989)) propose cutting the hierarchy. When the description of the object is similar enough to the description of the concept in which the object is classified, future descent (into subconcepts) is unnecessary, because enough information is already found in the concept description. In ADECLU(2 and /S), this possibility can be detected using the score function $sc(O,C,P)$. The rule is the following:

Let (C, P) be the pair which optimizes the suitability criterion $S(O, C, P)$. Thus ADECLU integrates the object O into the concept C (of the partition P). Afterwards, we:

"STOP DESCENT in the hierarchy IF" :

$$\frac{sc(O,C,P)}{f(|C|)} > \text{"cutoff" } \#C$$

where $\#C$ is the total number of characteristics (with $w \geq 0$) in C

The left term of this inequality has a high value when :

1) the number of characteristics (with $w_j \neq 0$) which match between the descriptions of the object and the concept (i.e. $j(O) \in V_j$) is high.
2) $\#_{w\neq 0}C$ (the number of characteristics in C with $w_j \neq 0$) is high
3) the characteristics of C are discriminant ($w_j \approx 1$).

The value of "$\#_{w\neq 0}C/\#C$" gives the percentage of characteristics which supply useful information to the concept description. If this percentage is small, few attributes are used at this level and it is thus necessary to pursue the specification of this concept. Thus the conditions 2) and 3) are satisfied when the concept description is *specific* and *contrasted* enough with respect to the other concepts. Examples (see (Decaestecker, 1990)) suggest taking a cutoff value of 0.5.

When the descent is stopped in *C*, the object *O* is integrated into *C*, but not into its sub-concepts. At a later stage in the incremental process, the definition of *C* and its location in the hierarchy can be changed. It would then be possible to continue (if necessary) the classification of *O* into the lower levels of the concept hierarchy.

5. PREDICTIVE POWER (with and without NOISE)

Recall two principal aims of Concept Formation :

- to condense (to summarise) information
- to use this information to predict unknown values

The evaluation procedure to measure the predictive power (ability to make predictions concerning unknown attributes) of the hierarchies produced by Concept Formation systems, is inspired from the methods of supervised learning. (For further details, see (Fisher, 1987)(Gennari & al, 1989)). The system incorporates a number of cases (the so-called training set) of a data base. A hierarchy of concepts is thus created and will be used later to classify the remaining cases (the so-called test set), which do not include the value of one attribute. The value of this attribute is inferred from the result of the classification, by predicting the value taken from the concept into which the test case was classified. One can then calculate the percentage of correct predictions as a function of the number of cases present in the training set.

To investigate the effects of noise (i.e here, the incorrect reporting of an attribute's value) on the performances of ADECLU, we begin by randomly replacing attribute values in all the data base, with a fixed probability (e.g 25%). Then ADECLU is submitted to the same experiments as described before.

We will perform experiments on 4 classes (4 x 17 cases of 50 attributes), extracted from a data base of soybean diseases (Michalski & al, 1980). TABLES 3 and 4 below present the percentages of good predictions for the attribute *CL* indicating the membership of one of the four classes, and the attribute *X8*, which is more difficult to predict in the presence of noise. The second data base studied has 140 descriptions of mushrooms classed according to degree of edibility. The third data base has 148 examples from the lymphography domain[1] with 4 possible final diagnostic classes. This data set was not submitted to a detailed checking and thus may contain errors in attribute values. TABLES 5 and 6 below present the percentages of good predictions for the attribute *CL* indicating the membership of one of the different classes.

These predictions have been made either with or without noise in the data. The experiments have been repeated several times on each data base presented in different orders. The tables before summarise the

[1] The data was obtained from the University Medical Centre, Institute of Oncology, Ljubljana, Yugoslavia.

performances obtained, indicating, for each version, the average of the results as a function of the cardinal of the "training set".

To better visualise the effect of the selection of attributes, an other version of ADECLU was introduced (called ADECLU1-2) where all the weights w_j of ADECLU2 are fixed at 1.0 (no selection of relevant characteristics). Note that this version gives worse results and its performance becomes mediocre in the presence of noise. ADECLU1 performs very well in the absence of noise, but its performance degrades in the presence of noise. This degradation increases with the growth of the training set (and therefor with the introduction of noise). The versions which behave the best, in the two cases, are ADECLU2 and ADECLU/S with a slight superiority of ADECLU/S for larger training sets (where "typicality" can be found and can play its role). The "statistical" selection of relevant attributes and of typical values, thus appears to be beneficial in its ability to predict unknown values.

card of training set:	5	10	15	20	25	30
WITHOUT NOISE						
ADECLU1:	88	87	90	92	94	91
ADECLU1-2:	94	78	84	81	88	83
ADECLU2:	91	91	92	91	92	91
ADECLU/S:	90	87	**98**	**97**	**95**	**97**
25% of NOISE						
ADECLU1:	85	80	81	78	80	76
ADECLU1-2:	85	55	56	56	55	54
ADECLU2:	85	85	87	78	79	80
ADECLU/S:	85	74	76	**80**	**80**	**82**

TABLE 3 : % (average) of good predictions
Data base "Soya", attribute *CL*

card of training set:	5	10	15	20	25	30
WITHOUT NOISE						
ADECLU1:	83	92	93	94	92	91
ADECLU1-2:	83	88	81	83	84	72
ADECLU2:	84	89	95	90	97	95
ADECLU/S:	83	90	91	**92**	**98**	**97**
25% of NOISE						
ADECLU1:	55	64	60	60	57	60
ADECLU1-2:	55	61	62	51	51	49
ADECLU2:	54	62	63	70	61	68
ADECLU/S:	54	62	62	67	**65**	**69**

TABLE 4 : % (average) of good predictions
Data base "Soya", attribute *X8*

card of training set:	20	40	60	80	100
WITHOUT NOISE					
ADECLU1:	63	78	84	74	80
ADECLU1-2:	58	56	58	54	54
ADECLU2:	68	82	79	89	87
ADECLU/S:	72	77	75	80	**88**
25% of NOISE					
ADECLU1:	48	55	45	46	59
ADECLU1-2:	35	41	25	38	37
ADECLU2:	48	60	65	69	**69**
ADECLU/S:	53	59	60	62	**69**

TABLE 5 : % (average) of good predictions
Data base "Mushroom", attribute *CL*

card of training set:	20	40	60	80	100
WITH NOISE					
ADECLU1:	66	63	69	66	69
ADECLU1-2:	61	56	67	60	67
ADECLU2:	67	76	73	71	**73,6**
ADECLU/S:	63	66	67	68	**74,3**

TABLE 6 : % (average) of good predictions
Data base "Lympho", attribute *CL*

6. CONCLUSIONS and FUTURE DIRECTIONS

A central idea in ADECLU is that of "contrast". It occurs in the definition of the evaluation criterion and in the representation of concepts. In fact, a concept is an entity which is defined not only by its own characteristics, but also by contrasting it with other concepts of the partition. Thus, to test the hypothesis that an object O must be attributed to a concept C, the Suitability Criterion uses :

in a *positive* fashion : – the similarity between the description of O and C
– the dissimilarity between the descriptions of O and the other concepts
and in a *negative* fashion : – the similarity between the descriptions O and the other concepts
– the dissimilarity between the descriptions of O and C.

Each of the characteristics of a concept contributes to the calculation of the score as a function of its relevance (w_j) in the concept description. In ADECLU2 and /S, this relevance will be all the greater when the characteristic is discriminating for the concept. In ADECLU/S, the selection of typical values (from the values observed in the concept) occurs with the aim of increasing the relevance of characteristics. It increases the contrast between concept descriptions, by extracting from the typical set Vj (and transfering to ${}^{c}Vj$) the values which are too frequently observed in the other concepts. This typicality is thus of a statistical nature and is "acquired' from the data provided to ADECLU. To learn it correctly, it is necessary to have a sufficient number of objects. If not, ADECLU2 is preferable.

After experimentation, it appears that the use of "contrasted" descriptions among concepts is beneficial not only for the quality of data classification (cf (Decaestecker, 1990)), but equally for the predictive power of the conceptual hierarchies, especially in the presence of noise.

Other definitions of typicality can be envisaged (cf (Lebbe & al, 1988)). ADECLU/S uses a "strong" typicality ("frequency in C" > "frequency in $P\backslash\{C\}$"), which generalises the notion of "discriminant value" ("frequency in $P\backslash\{C\}$" = 0). This definition of typicality implies that a value be selected in at most one concept. Studies are underway on the different ways of dropping this constraint, aiming at a less draconian selection of values. Initial results are encouraging.

The work on ADECLU can be related to other Conceptual clustering methods. For example, WITT (Hanson & al. 86), a non incremental system, is also related to the notion of "concept contrasting", mentioned above. The concept representation also uses contingency tables but one for each pair of attributes. This necessitates considerable storage costs ($A[A-1]/2$ tables for A attributes - too high for incremental processes). ADECLU(2 and /S) each requires only A (2x2) contingency tables with easy updating.

COBWEB (Fisher, 1987) and CLASSIT (Gennary & al, 1989) are similar to ADECLU in their clustering processes (hill climbing search with operators), but differ in their concept representations and do not select attribute and typical values. Hence, COBWEB and ADECLU(2 and /S) store the same basic knowledge : the frequencies of observed values in each concept, but ADECLU uses this knowledge to select relevant characteristics to classify new objects and to predict unknown values. CLASSIT2 (Gennari, 1989) is an extension of CLASSIT that includes a mechanism to focus attention upon a subset of attributes that are more "salient". Attributes are inspected in sequence, where the inspection order is determined by a "salience" measure, which is the contribution of one attribute to his evaluation measure (category utility). Attribute inspection stops when the remaining attributes cannot change the clustering decision.

In (Fisher, 1989), the author presents a pruning method of hierarchy (for noise-tolerant Conceptual Clustering), which uses past-performance (number of times the attribute was correctly predicted during training). During test, classification is stopped in a node that has historically outperformed its descendants (in terms of predicting missing attribute). We have compared this method with our cutoff method in ADECLU. First experiments (Decaestecker, 1991) show that Fisher method does not ameliorate our results in noisy domains.

REFERENCES

Anderberg M.R. "Cluster analysis for applications", Academic Press, New York, 1973.

Decaestecker C. "Formation incrémentale de concepts par un critère d'adéquation". Proceedings of "4èmes Journées Française d'Apprentissage", 1989 (a).

Decaestecker C. "Incremental Concept Formation with Attribute Selection", Proceedings of the 4th EWSL (European Working Session on Learning), pp 49-58, 1989 (b).

Decaestecker C. "Description Contrasting in ADECLU : ADECLU/S", Technical Report IA-01-90, CADEPS (CAIRU), Université Libre de Bruxelles, Belgium, 1990.

Decaestecker C. Thesis, in preparation, Université Libre de Bruxelles, Belgium, 1991.

Fisher D.H. "Knowledge acquisition via incremental conceptual clustering", *Machine Learning,* 2, pp 139-172, 1987.

Fisher D.H. "Noise-Tolerant Conceptual Clustering", Procceedings of IJCAI-89, pp 825-830, 1989.

Gennari J.H., Langley P. and Fisher D. "Models of Incremental Concept Formation", *Artificial Intelligence,* vol. 40, No 1-3, 1989.

Gennari J.H. "Focused Concept Formation", Proceedings of the 6th International Workshop on Machine Learning, pp 379-382, 1989.

Goodman L.A. and Kruskal W.H. "Measures of association for cross classifications", *J.of the American Stat. Assoc.,* vol. 49, pp 732-764, 1954.

Hanson S.J and Bauer M. "Machine Learning, Clustering and Polymorphy". In L.N. Kanal & J.F. Lemmer (Eds), *Uncertainty in Artificial Intelligence,* North-Holland, 1986.

Lebbe J., Vignes R., Darmoni S. "Les objets symboliques qualifiés, Application au diagnostic médical", Proceedings of the "2èmes Journées Symbolique-Numérique", Report n° 487, LRI, Université Paris-Sud (Orsay), 1988.

Lebowitz M. "Experiments with incremental concept formation : UNIMEM", *Machine Learning,* 2, pp 103-138, 1987.

Michalski R.S. and Chilausky R.L. "Learning by being told and learning from examples ...", *Policy analysis and Information Systems,* vol.4, No 2, June 1980.

Quinlan J.R. "Induction of decision trees", *Machine Learning,* 1, pp 81-106, 1986.