A RECENT ADVANCE IN DATA ANALYSIS: CLUSTERING OBJECTS INTO CLASSES CHARACTERISED BY CONJUNCTIVE CONCEPTS

by

Ryszard S. Michalski
Robert E. Stepp
Edwin Diday

Invited chapter in the book Progress in Pattern Recognition, Vol. 1, L. Kanal and A. Rosenfield (Editors), 1981.

REPRINTED FROM:

PROGRESS IN PATTERN RECOGNITION

Volume 1

edited by

Laveen N. KANAL

and

Azriel ROSENFELD

Department of Computer Science
University of Maryland, College Park, Md.

NORTH-HOLLAND PUBLISHING COMPANY – AMSTERDAM • NEW YORK • OXFORD

PROGRESS IN PATTERN RECOGNITION
L.N. Kanal and A. Rosenfeld (editors)
North-Holland Publishing Company, 1981

A RECENT ADVANCE IN DATA ANALYSIS: Clustering Objects into Classes Characterized by Conjunctive Concepts

Ryszard S. Michalski
Robert E. Stepp

Department of Computer Science
University of Illinois
Urbana, Illinois
U.S.A.

Edwin Diday

Institut National de Recherche
en Informatique et en Automatique
Domaine de Voluceau, Rocquencourt
France

Clustering is described as a multistep process in which some of the steps are performed by a data analyst and some by a computer program. At present, those performed by a computer program do not produce any description of the generated clusters. The recently introduced method of *conjunctive conceptual clustering* overcomes this problem by requiring that each cluster has a conjunctive description built from relations on object attributes and closely "fitting" the cluster. The paper explains the above clustering method in terms of *dynamic clustering* and shows by an example its advantages over methods of numerical taxonomy from the viewpoint of cluster interpretation.

I. Introduction

Clustering is usually viewed as a process of partitioning data into groups of "similar" objects. Accepting this view, the thrust of research in the area of cluster analysis has been toward determining various object similarity or proximity measures, and developing clustering techniques utilizing them. A large number of such measures and corresponding clustering methods have been developed to date. Comprehensive surveys can be found in Sokal and Sneath [35], Dorofeyuk [13], Cormark [7], Zagoruiko [39], Anderberg [1], and Diday and Simon [9]. In all these methods, the task is to determine clusters, such that objects in the same cluster exhibit a high degree of similarity (high intra-cluster similarity), while objects from different clusters exhibit a low degree of similarity (low inter-cluster similarity).

The above approach to clustering has some major drawbacks. A researcher analyzing data typically wants not only to determine clusters but, even more, wants to know the underlying conceptual meaning behind them. The conventional methods, however, leave the subject of interpreting the clusters to the researcher. Moreover, clusters obtained merely on the basis of object similarity may not have any simple conceptual interpretation. This is so because the clusters are obtained without taking into consideration any linguistic constructs or concepts which people use in characterizing groups of objects. To overcome this problem, the method of conceptual clustering has recently been introduced and implemented (Michalski [25], Michalski and Stepp [26,27]).

In conceptual clustering, clusters are assumed to be not just collections of similar objects, but groups of objects that have a simple conceptual interpretation. Such an interpretation may be just a name characterizing a collection of objects in a certain configuration (e.g., ring-shaped, U-shaped, T-formation) or a description characterizing properties of the configuration. In the method presented by Michalski and Stepp [25,26], the descriptions are conjunctive statements built from relations on selected object attributes. In this paper, after examining various components of the clustering process, we

explain conjunctive conceptual clustering in the context of the so-called dynamic clustering method (Diday et al, [8,12]). An experiment then is described which contrasts conjunctive conceptual clustering with several methods of numerical taxonomy.

II. An Overview of Clustering Problems

A. Problem Classes

From the viewpoint of applications, it is useful to distinguish three classes of clustering problems on the basis of the dimensionality of objects to be clustered:

1. One-dimensional clustering (quantization of variables)

Suppose that data describing objects involve continuous variables or discrete variables with ranges of values that are significantly larger than necessary for a given problem. In such cases one wants to reduce the number of distinct values of the variables. This can be done by identifying equivalence classes of the values (i.e., clusters of values) which should be treated as single units. For example, in image processing, the scanners usually distinguish between a large number of gray levels but only a few levels may be needed for solving a particular image recognition problem. Rosenfeld [34] and Pratt [31] have shown that clustering methods can be used for making such a reduction. Nubuyaki [30] proposed a clustering algorithm for this purpose in which the clustering optimality criterion was the sum of the squares of intra-cluster distances. Several one-dimensional clustering techiques have been used in the LANDSAT system to reduce the range of intensity values in spectral components of earth satellite images. One of them, described by Swain [36], applies ISODATA methods [3], and another, described by Roche [32], uses dynamic clustering. One-dimensional clustering techniques have also been proposed for general data reduction problems (Lowitz [19]).

2. Two-dimensional clustering (segmentation)

This type of clustering occurs most often in image processing, where one searches for segments of an image in which all picture elements share some common properties, e.g., have similar gray level or similar texture. Coleman [6] defined region segmentation as a problem of non-supervised learning (clustering), and applied to it the k-means algorithm (MacQueen [20]). Do-Tu and Installe [14] used the ISODATA method to segment LANDSAT data. Other clustering algorithms for image segmentation have been proposed by Fukada [15], Kasvand [17], and Backer [2]. Yokoya et al. [38] constructed a partition of an image depending on two parameters and showed that by varying those parameters, a hierarchy of clusters can be obtained. Clustering can also been used to extract contours. For example, Haralick and Shapiro [16] used clustering for decomposing images into polygonal contours, and Charles and Lechevallier [4] used it for polynomial approximation of contours in cartography.

3. Multi-dimensional clustering

In multi-dimensional clustering, objects are partitioned into clusters in a description space spanned by many attributes characterizing the objects. The basis for clustering is typically a "similarity" measure between objects, which is defined as a function of the object attributes. Clusters are determined as collections of objects with high intra-cluster similarity and low inter-cluster similarity.

In this paper we are concerned with multi-dimensional clustering.

B. Components of a Clustering Problem Specification

To formulate a clustering problem the data analyst must specify the following basic method-independent components:

(1) the set of objects to be clustered,

(2) the set of attributes (variables) to be used in describing objects,

(3) the method of encoding variables, including the specification for each variable of its domain and its measurement scale,

(4) a principle for grouping objects into clusters (traditionally, a measure of object similarity; in this paper, the membership of an object in a class characterized by a conjunctive concept),

(5) the inter-cluster structure, which defines the desired relationship among clusters, e.g., whether the clusters should be disjoint, overlapping, or hierarchically organized sets.

Each of these components is described in detail below. An illustration of these components is given in an example in section VI.

(1) and (2) Objects to be clustered and the choice of attributes

Typically, objects to be clustered come from an experimental study of some phenomenon. One basic property they all possess is that they are describeable by the same set of attributes (variables). These attributes, however, are not always all relevant to the clustering problem. The task of detecting the relevant attributes can be considered as either a separate problem or as an integral part of clustering. In conventional approaches, the selection of relevant attributes is treated as a separate preliminary step. In conjunctive conceptual clustering (sections IV and V of this paper), the selection of relevant attributes is an integral part of the whole method and is performed simultaneously with the formation of clusters.

(3) The encoding of attributes

Attributes represent distinct measurements or observable properties. In the case of physical objects attributes may be, for example, size, weight, temperature, color, shape, chemical structure, etc. The initial encoding of the attributes is dictated by the measurement devices used, or by an established convention. The attributes may be measured on different scales, such as nominal, ordinal, interval, ratio, and absolute. In a simple case, one can only distinguish between qualitative attributes (the nominal scale) and quantitative attributes (the remaining scales). These initial measurements are subject to a problem-dependent transformation, which may reduce the precision of the quantitative attributes or replace subranges of their values by qualitative properties (e.g., a numerical size may be replaced by characterizations such as "small," "medium," or "large").

The subject of optimal reduction of variable precision was recently considered by Taleng [37], who describes an adaptive technique of variable encoding. Other methods of transforming the quantitative attributes into qualitative ones are described by, for example, Anderberg [1] and Lechevallier [18]. Zagoruiko [40] discusses a method of optimal quantization by using the concept of mutual information.

(4) A principle for grouping objects into clusters

a. The measure of similarity

The traditional principle for grouping objects into clusters utilizes some measure of object similarity, usually a reciprocal of a distance measure. Among the many distance measures proposed for clustering (Diday and Simon [9]) one can distinguish between quantitative measures (Figure 1) and qualitative (binary) measures (Figure 2). The task of selecting a measure of similarity for a given clustering problem has been studied by Rohlf [33], Chernoff [5], and Diday and Govaert [8].

b. Membership in a class characterized by a single concept

In conceptual clustering in general [25], objects are assembled into clusters that represent single human concepts (linguistic terms or simple logical functions defined on such terms). In the conjunctive conceptual clustering the concepts are logical products of relations on selected object attributes.

(5) The inter-cluster structure

Let E be a set of objects to be clustered and E_1, E_2,..., E_k clusters into which E is partitioned. Let $\alpha(E_i) = \alpha_i$ denote a description of cluster E_i (a conjunctive statement covering E_i, see section IV). In general, a description α_i is satisfied not only by all the observed objects in E_i, but also by some unobserved objects. Based on the relationships among the clusters E_i, i=1,2,...,k, or among the cluster descriptions α_i, four different types of inter-cluster structures are commonly distinguished in the literature:

- The partition structure: a set of clusters whose union is the set E, and whose descriptions are all disjoint[1] (this implies that the clusters themselves are disjoint),
- The overlapping structure: a set of clusters that includes at least one intersecting pair. When some descriptions intersect but corresponding clusters do not (i.e., the intersection of these descriptions contains only the unobserved events), the structure is called weakly overlapping, otherwise it is called strongly overlapping.
- The hierarchical structure: the first level clusters represent a partition structure of the whole set E; clusters at a lower level are elements of partition structures of the corresponding clusters one level higher,
- The bipolar structure: a partition structure that consists of pairs of clusters with maximally contrasting representations.

These structures are illustrated in Figure 3.

III. The Dynamic Clustering Method

The dynamic clustering method (Diday et al, [8-12]) is a class of clustering techniques that find clusters iteratively by alternately applying a representation function and an allocation function (explained below) until a local optimum of the assumed criterion of clustering optimality is achieved. The earlier algorithms such as "k-means" (MacQueen [20]) and the "center adjustment algorithm" (Meisel [28]) are special cases of dynamic clustering, in which the description space is Euclidean and the cluster representation is the cluster mean.

(a) $$d(X_p,X_q) = \left[\sum_{i=1}^{n} |x_{pi}-x_{qi}|^{\frac{1}{\lambda}}\right]^{\lambda}$$ Minkowsky

(b) $$d(X_p,X_q) = \sum_{i=1}^{n} \frac{|x_{pi}-x_{qi}|}{|x_{pi}+x_{qi}|}$$ Camberra

(c) $$d(X_p,X_q) = \max_{i=1}^{n} |x_{pi}-x_{qi}|$$ Chebyshev

(d) $$d(X_p,X_q) = (X_p-X_q)^T W(X_p-X_q)$$ Quadratic

$$= \sum_{j=1}^{n} \left(\sum_{i=1}^{n} (x_{pi}-x_{qi})w_{ji} \right) (x_{pj}-x_{qj})$$

(e) $$d(X_p,X_q) = \frac{\sum_{i=1}^{n} (x_{pi}-\bar{x}_i)\,(x_{qi}-\bar{x}_i)}{\left[\sum_{i=1}^{n} (x_{pi}-\bar{x}_i)^2 \sum_{i=1}^{n} (x_{qi}-\bar{x}_i)^2\right]^{1/2}}$$ Correlation

(f) $$d(X_p,X_q) = \sum_{i=1}^{n} w_i |x_{pi}-x_{qi}|$$ "City block"

(g) $$d(X_p,X_q) = \sum_{i=1}^{n} \left[\sum_{r=1}^{m} x_{ri}\right]^{-1} \left[\frac{x_{pi}}{\sum_{j=1}^{n} x_{pj}} - \frac{x_{qi}}{\sum_{j=1}^{n} x_{qj}}\right]^2$$ Chi-square

(h) $$d(X_p,X_q) = 1 - \frac{2}{n(n-1)} \sum_{i=1}^{n} \sum_{j=1}^{i-1} \Gamma(x_{qi}-x_{qj})\,\Gamma(x_{pi}-x_{pj})$$ Kendall's Rank Correlation

where $d(X_p,X_q)$ - the distance between objects represented by attribute vectors X_p and X_q ($p,q \in \{1,\ldots,m\}$, m-number of objects)

x_{ij} - the value of attribute j in X_i

W - problem-specific weight values (in eq.(f) W is an n-element vector; in eq.(d) W is an n x n matrix)

$\Gamma(x)$ - the sign of x (-1, 0, or 1, when $x<0$, $x=0$, or $x>0$, respectively)

Figure 1. Some Quantitative Distance Measures

(a)	$d(X_p,X_q) = \frac{a}{a+b+c+e}$	Russel and Rao
(b)	$d(X_p,X_q) = \frac{a}{a+b+c}$	Jaccard and Needham
(c)	$d(X_p,X_q) = \frac{a}{2a+b+c}$	Dice
(d)	$d(X_p,X_q) = \frac{a}{a+2(b+c)}$	Sokal and Sneath
(e)	$d(X_p,X_q) = \frac{a+e}{a+b+c+e}$	Sokal and Michener
(f)	$d(X_p,X_q) = \frac{a}{b+c}$	Kulzinsky
(g)	$d(X_p,X_q) = \frac{a+e}{a+e+2(b+c)}$	Rogers and Tanimoto
(h)	$d(X_p,X_q) = \frac{ae-bc}{ae+bc}$	Yule
(i)	$d(X_p,X_q) = \frac{ae+bc}{[(a+b)(c+e)(a+c)(b+e)]^{1/2}}$	correlation

where a, b, c, and e are defined as:

$$a = \sum_{i=1}^{n} x_{pi} \cdot x_{qi} \qquad c = \sum_{i=1}^{n} x_{pi} \cdot (1-x_{qi})$$

$$b = \sum_{i=1}^{n} x_{qi} \cdot (1-x_{pi}) \qquad e = \sum_{i=1}^{n} (1-x_{pi}) \cdot (1-x_{qi})$$

Some Qualitative Binary Distance Measures
Figure 2

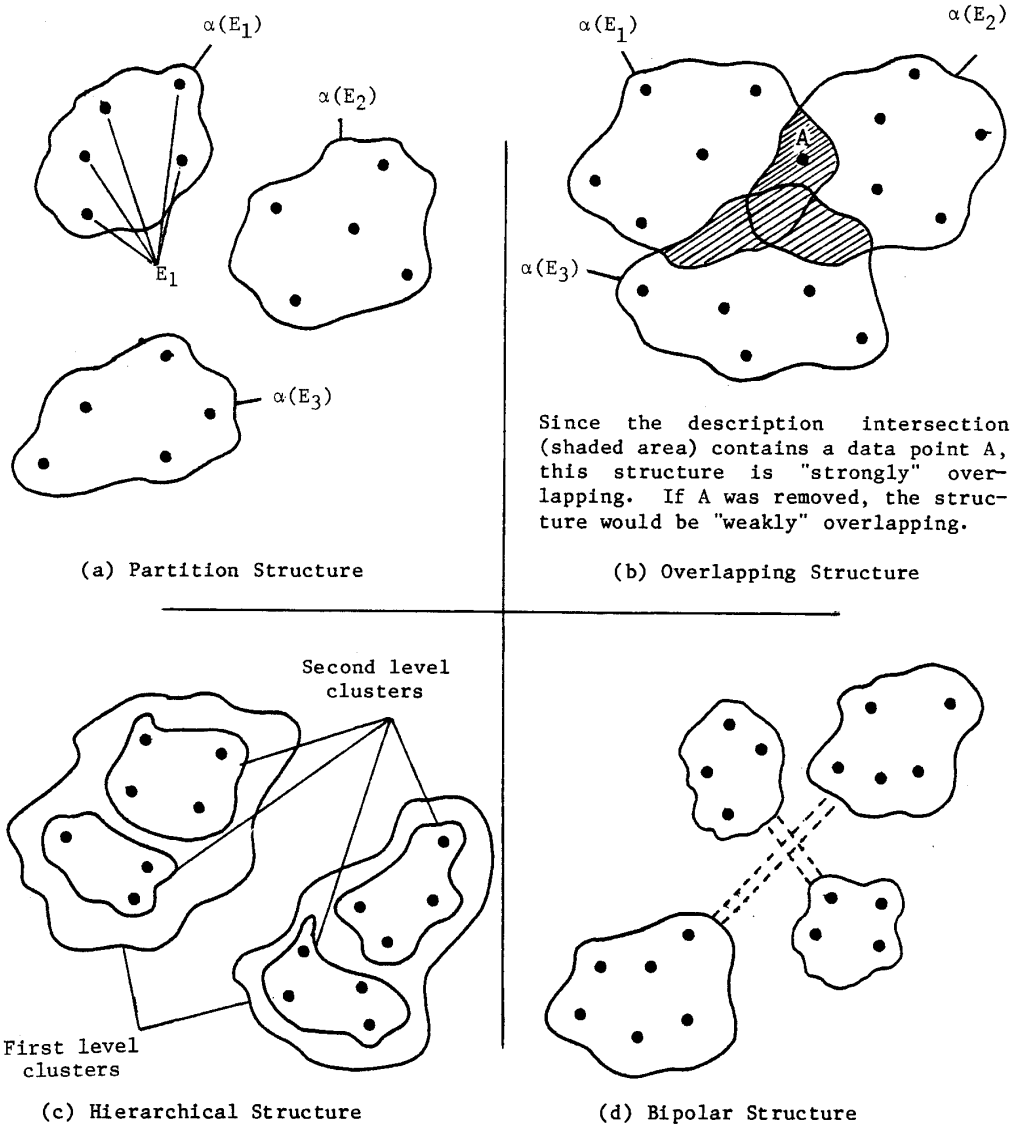


Figure 3. Examples of Inter-cluster Structures

The dynamic clustering method can be viewed as a general framework that is particularized to a specific technique by specifying appropriate problem-dependent components (in addition to the five basic components described earlier in section IIB). These additional method-dependent components include:

(6) the cluster representation scheme, which defines the means for formally representing clusters,

(7) the representation function, which, given the objects in a cluster, determines their representation,

(8) the allocation function, which, given a cluster representation, determines observed objects belonging to the cluster,

(9) the clustering optimality criterion, which defines the optimal clustering of objects (ideally, such a measure should take into consideration both the measure of fit of clusters to cluster representations and the utlility of the clusters).

Each of these items will now be discussed in detail.

(6) Cluster representation schemes

A cluster representation is a mathematical or geometrical construction that simply and generally characterizes objects in the cluster. An elementary cluster representation scheme is to select one or a few sample objects from the cluster. Variations on this theme include selecting the object that corresponds to an event at the center of mass of the cluster, or the object closest to it (when no observed event is at the center of mass). An alternative scheme is to select the set of "most outstanding representatives" of the cluster, defined as a set of the r most different objects found in the cluster (where r is a constant). Such a set can be determined by the event selection program ESEL (described by Michalski and Larson [24]). Figures 4a and 4b illustrate the above two schemes of cluster representation. Other schemes include representing a cluster by the line of least inertia (Figure 4c), a normal distribution (Figure 4d), a node in a classification tree (Figure 4e), and by a conjunctive statement (Figure 4f), which is the representation used in this paper.

(7) The representation function

Given a representation scheme and a set of clusters of objects, the representation function determines the "best" representation for the clusters under the assumed scheme. Algorithms for computing the representation function depend strongly on the choice of the scheme for cluster representation. Formally, the representation function is a mapping

$$g : \{C^k\} \rightarrow \{L^k\} \qquad (1)$$

where $\{C^k\}$ is a set of clusterings (each clustering is a collection of k clusters),
$\{L^k\}$ is a set of clustering representations (each clustering representation is a collection of representations of individual clusters).

(8) The allocation function

The allocation function is the inverse of the representation function: given a cluster representation, it determines the objects that belong to each cluster. Formally, it is a mapping

$$f : \{L^k\} \rightarrow \{C^k\} \qquad (2)$$

(9) The clustering optimality criterion

One way of measuring the optimality of clustering is to measure the "fit" between clusters and cluster representations. Such a measure of fit is usually defined as the sum of the degrees of fit between each object in a cluster and the cluster representation, and thus is additive (this is not the case in conceptual clustering, since it tries to capture "Gestalt properties" of clusters). The

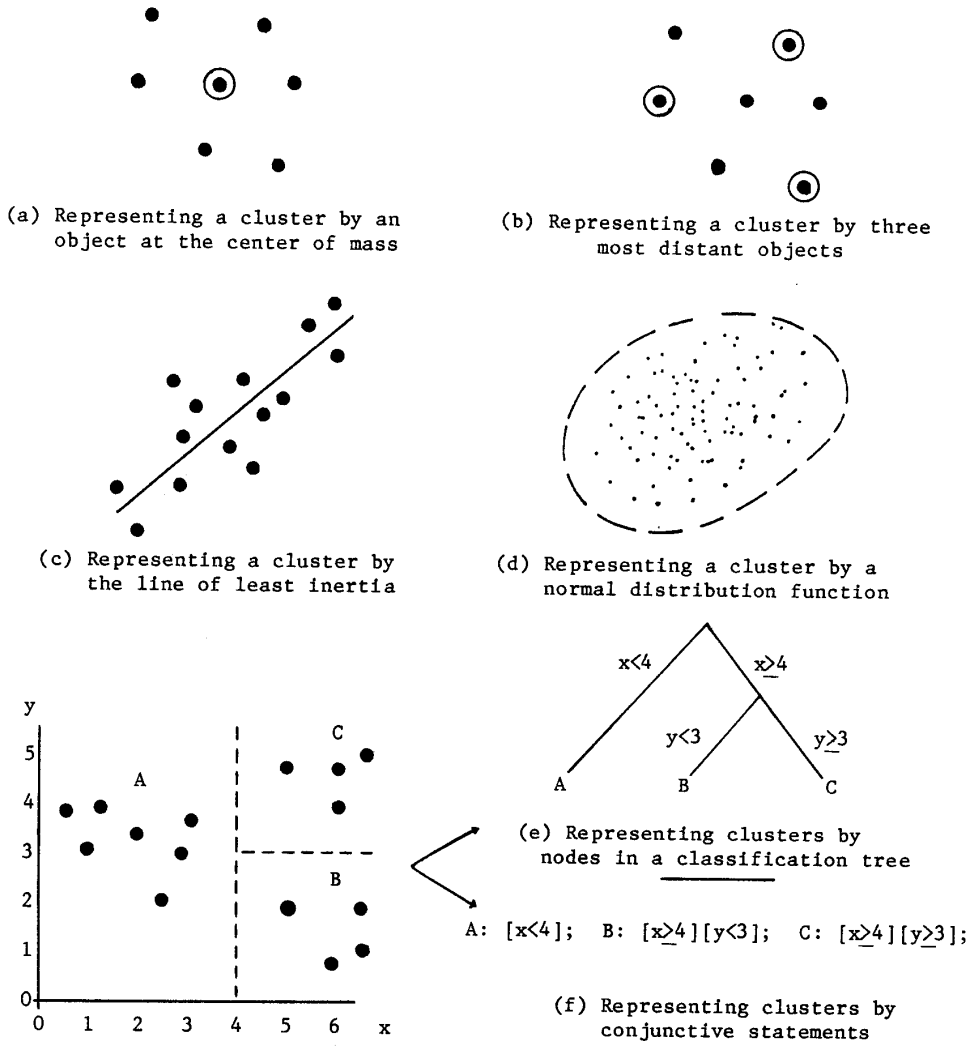


Figure 4. Examples of Cluster Representation Schemes

measure of fit can be stated formally as a mapping

$$DF : \{C^k\} \times \{L^k\} \rightarrow [0,1] \qquad (3)$$

where x denotes the cartesian product and [0,1] is the interval of possible values of the degree of fit.

In evaluating the clustering optimality one may also consider the utility of the clusters.

In such cases, the clustering optimality criterion would be a combined measure, e.g., a linear function with weights representing the relative importance of the degree of fit and the utility:

$$W(C,L) = w_1 \cdot DF(C,L) + w_2 \cdot U(C,L) \qquad (4)$$

where $U(C,L) \in [0,1]$ is a function that measures the utility of representing cluster C by representation L. An alternative way of combining the above measures of clustering optimality is to use the "lexicographical functional with tolerances" described in section IV. Section IV also describes some other measures for characterizing clustering optimality.

The Control Structure of the Dynamic Clustering Algorithm

Given a set of objects E, and integer k, the dynamic clustering method partitions E into k clusters that are locally optimal according to the assumed criterion of clustering optimality, W(C,L). Beginning with some initial representation of the k clusters, chosen randomly, a sequence of iterations is performed, each consisting of (a) finding the clusters that best fit the given cluster representations obtained so far, and (b) choosing the representations that best fit the obtained clusters. During this process, the clustering optimality criterion is monitored and when improvement ceases, the process terminates.

Let $C_0, C_1, C_2, \ldots$ and $L_0, L_1, L_2, \ldots$ be sequences of clusterings and representations, respectively, obtained in iterations $0,1,2,\ldots$. If the sequence of values of the clustering optimality criterion $W(C_0,L_0)$, $W(C_1,L_1)$, $W(C_2,L_2)$, ... is guaranteed not to decrease, then it must converge to some (local) maximum value. The convergence properties of W and the type of similarity function that guarantees the convergence are discussed in Diday et al [11].

IV. Conjunctive Conceptual Clustering

The conjunctive conceptual clustering method, as described by Michalski and Stepp [26,27], can be viewed in the framework of dynamic clustering if one assumes that the cluster representation has a special form: a conjunctive concept closely "fitting" a collection of objects. This way of presenting conceptual clustering seems to be useful for the simplicity of theoretical presentation, especially for those acquainted with conventional clustering methods. It should be noted, however, that the actual computational techniques in conceptual clustering are quite different from anything previously applied under the dynamic clustering framework. These techniques employ the formulation and methodology developed for the optimization of variable-valued logic expressions (Michalski [21,22]).

Representation scheme

Conjunctive conceptual clustering uses two cluster representation schemes: (1) a preliminary scheme: a single object selected from a cluster (central or extreme), called the seed of the cluster, and (2) a final scheme: a conjunctive statement that describes objects in the cluster (Figure 4f). This conjunctive statement (called a VL_1 conjunctive predicate or a VL_1 complex) is an expression in the variable valued logic system VL_1 (Michalski [21,23]).

Suppose that $x_1, x_2, \ldots, x_n$ are variables selected to represent objects to be clustered. We will assume that each variable, x_i, $i \in \{1,2,\ldots,n\}$, has an assigned domain, $D(x_i)$, that specifies all possible values the variable can take for any object in the collection to be clustered. The number of such values is given by d_i. The domains (after final transformation) are assumed to be finite, and represented generally as $D(x_i) = \{0,1,2,\ldots,d_i-1\}$. We distinguish between nominal variables (or qualitative measurements), whose domains are unordered sets, linear variables (or quantitative measurements) whose domains are linearly ordered sets, and structured variables, whose domains are tree-ordered sets. An example

of a nominal variable is "color" or "blood type;" examples of linear variables are rank, size, or quantity of something; an example of a structured variable is shape, whose values may be triangle, rectangle, pentagon, ..., or polygon which represents a more general concept (a parent node in the tree-structured domain). For simplicity, we assume that variables are either nominal or interval.

The description space spanned over variables $x_1, x_2, \ldots, x_n$ is called the event space. Each point in this space (event) is a vector of specific values of variables $x_1, x_2, \ldots, x_n$. An event that is a description of some object in the collection to be clustered is called an observed event. Other events are called unobserved events. A VL_1 conjunctive predicate or VL_1 complex (briefly, ℓ-complex) is a logical product of relational predicates or selectors, defined as forms:

$$[x_i \ \# \ R_i] \tag{5}$$

where R_i (the reference) is a list of values from the domain of variable x_i, i.e., $D(x_i)$,

\# (the relation) stands for the relational operator = (equal) or $\neq$ (not equal).

A selector $[x_i = R_i]$ (or $[x_i \neq R_i]$) is satisfied if the value of x_i is (is not) in relation = ($\neq$) with any (all) values in the set R_i. In the set-theoretic sense,

$[x_i = R_i]$ is equivalent to "value of $x_i \in \{R_i\}$" and
$[x_i \neq R_i]$ is equivalent to "value of $x_i \notin \{R_i\}$"

For example, the selector [length=small,medium] (value of length $\in$ {small,medium}) is satisfied whenever length has the value small or the value medium. The selector [length$\neq$medium] is satisfied by any value of length except medium. The notation of a selector may be simplified by using the "or" operator for linking values of nominal variables on the list R_i and using operators $< \ > \ \leq \ \geq$ and the range operator ".." in selectors with linear variables, as illustrated below. A set of objects that satisfy each selector in a complex is called an s-complex (set-complex). Thus, an ℓ-complex can be viewed as a description of an s-complex. The ℓ-complex:

$$[\text{height=tall}][\text{color=blue or red}][\text{length}\geq 2][\text{size}\neq\text{medium}][\text{weight=2..5}] \tag{6}$$

(the operation AND is implied by the concatenation of selectors) describes those objects that are tall, blue or red, with length ≥ 2, not medium size, and of weight 2 through 5. The set of all such objects constitutes the corresponding s-complex. The distinction between ℓ- and s- complexes is used to permit the application of logical or set-theoretic operators, respectively, whichever is more convenient. When this distinction is unimportant, the term complex will be used (without a prefix).

Not every collection of objects constitutes an s-complex, i.e., not every collection can be precisely described by an ℓ-complex. It is, however, possible to describe every collection of objects by a complex, if the complex is allowed to describe some additional objects (i.e., if it is permitted to be a generalized description of the collection). For example, events:

e_1: (blue, large, round)
e_2: (red, medium, round)

can be described by the complex: [color=blue or red][size$\geq$medium][shape=round]. This complex also covers the events:

e_3: (red, large, round)
e_4: (blue, medium, round)

which are distinct from e_1 and e_2. The number of such unobserved events contained in a complex is called the (absolute) sparseness of the complex.

Representation function (g)

In conceptual clustering, the representation function g is implemented as a procedure which, given a clustering C^k (a set of k clusters), selects k seeds $e_1, e_2, \ldots, e_k$, one from each cluster, and then determines a set of k disjoint complexes, $\alpha_1, \alpha_2, \ldots, \alpha_k$, such that

(1) complex α_i covers (contains) seed e_i,
(2) the union of complexes covers the set to be clustered E, and
(3) all k complexes together maximize the clustering optimality criterion.

This procedure is computationally very complex. The selection of seeds is initially done randomly, and then follows certain rules. These rules and the underlying algorithm are briefly described in section V.

Allocation function (f)

In contrast to function g, function f in conjunctive conceptual clustering is very simple. It is implemented as a procedure which, given a representation consisting of k complexes $\alpha_1, \alpha_2, \ldots, \alpha_k$, forms a clustering $C^k = \{E_1, E_2, \ldots, E_k\}$, where the cluster E_i contains observed events in complex α_i.

Clustering optimality criterion

The clustering optimality criterion specifies the desired properties of clusterings. The implemented method permits the user to maximize simultaneously one or more measures (elementary criteria) such as:

- the fit between the clusters and the data,
- the inter-cluster differences,
- the essential dimensionality,
- the simplicity of cluster representations.

The fit between the clusters and the data is computed as the negative of the total sparseness of the complexes defining clusters (i.e., the negative of the total number of unobserved events contained in the complexes). As the number of unobserved events in a complex decreases, the degree of overgeneralization of the complex decreases, and therefore the complex "fits" the observed events better.

Inter-cluster difference is measured by the sum of the degrees of disjointness between every pair of complexes representing clusters. The degree of disjointness of a pair of complexes is the number of selectors in both complexes after removing pairs of selectors that involve the same variable and intersect. For example, the pair of complexes

- [color=red][size=small or medium][shape=circle]
- [color=blue][size=medium or large]

has the degree of disjointness 3, because 2 of the 5 selectors intersect, namely [size=small or medium] intersects [size=medium or large]. Maximizing this criterion promotes clusters whose descriptions involve long sequences of different attribute values.

Essential (discriminative) dimensionality is defined as the number of variables that singly discriminate between all the clusters, i.e., which have different values in every cluster description (ℓ-complex). Single relations involving such variables are sufficient for distinguishing one cluster from the other clusters.

Simplicity of cluster representations is measured by the negative of the total number of selectors in all descriptions.

The above elementary criteria can be combined together into one general measure of clustering optimality through the use of the "lexicographical functional with tolerances" (LEF) [25]. The LEF is defined by a sequence of "criterion-tolerance" pairs (c_1,Δ_1), (c_2,Δ_2), ..., where c_i is a criterion (from the above list) and Δ_i is a "tolerance threshold" ($\Delta \varepsilon [0..100\%]$). In the first step, all clusterings are evaluated on the first criterion, c_1, and those that score best or within the range defined by the threshold Δ_1 from the best are retained. Next, the retained clusterings are evaluated on criterion c_2 and trimmed similarly as above using Δ_2. This process continues until either the subset of retained clusterings is reduced to a singleton (the "best" clustering), or the sequence of criterion-tolerance pairs is exhausted. In the latter case, the retained set contains clusterings that are considered to be equivalent with respect to the assumed criterion of optimality.

V. The Algorithm PAF

The preceding sections have described the conjunctive conceptual clustering method in the general framework of dynamic clustering. This section will present briefly the actual clustering algorithm PAF implemented as the inner part of the conceptual clustering program CLUSTER/PAF [26,27]. The outer program invokes the inner portion in a sequence of iterative steps to determine the "best" number of clusters and then recursively repeats the whole process to construct the next level of the cluster hierarchy [27].

The implemented algorithm, which reflects the dynamic clustering framework described in section III, proceeds as follows:

1. k events ("seeds") are selected from E. The seeds may be chosen randomly or according to some criterion,

2. For each seed, a set (star) of all maximally general complexes (i.e., with maximum sparseness) that cover this seed and do not cover other seeds is determined,

3. Complexes in stars are reduced by removing from selector references all unnecessary values, i.e., the values without which the complex still covers the same observed events,

4. From each (so modified) star, one complex is selected in such a manner that the obtained complexes are mutually disjoint, together cover all the data points, and optimize the given criterion of clustering optimality. The search strategy used to find such a collection of complexes is based on the A* search algorithm, developed in artificial intelligence (Nilsson [29]).

5. From each complex in the collection a new seed is selected and a new iteration of the algorithm begins. Two seed selection techniques are used. Seeds may be either central events, having the maximum number of properties in common with other observed events in the complex, or they may be borderline events, having the minimum number of properties shared. Central events are chosen as seeds as long as the clusterings improve with each iteration. When the improvement ceases, borderline events are selected.

6. The obtained clustering is evaluated using measures defined in the criterion of optimality (selected from: fit, inter-cluster differences, essential

dimensionality, or simplicity). If this is the first iteration, the clustering is stored, otherwise it is stored only if it is better than the previously stored one. This way, the optimality criterion is guaranteed not to decrease. The algorithm terminates when a specified number of iterations does not produce a better clustering.

Figure 5 shows the flow diagram summarizing these steps.

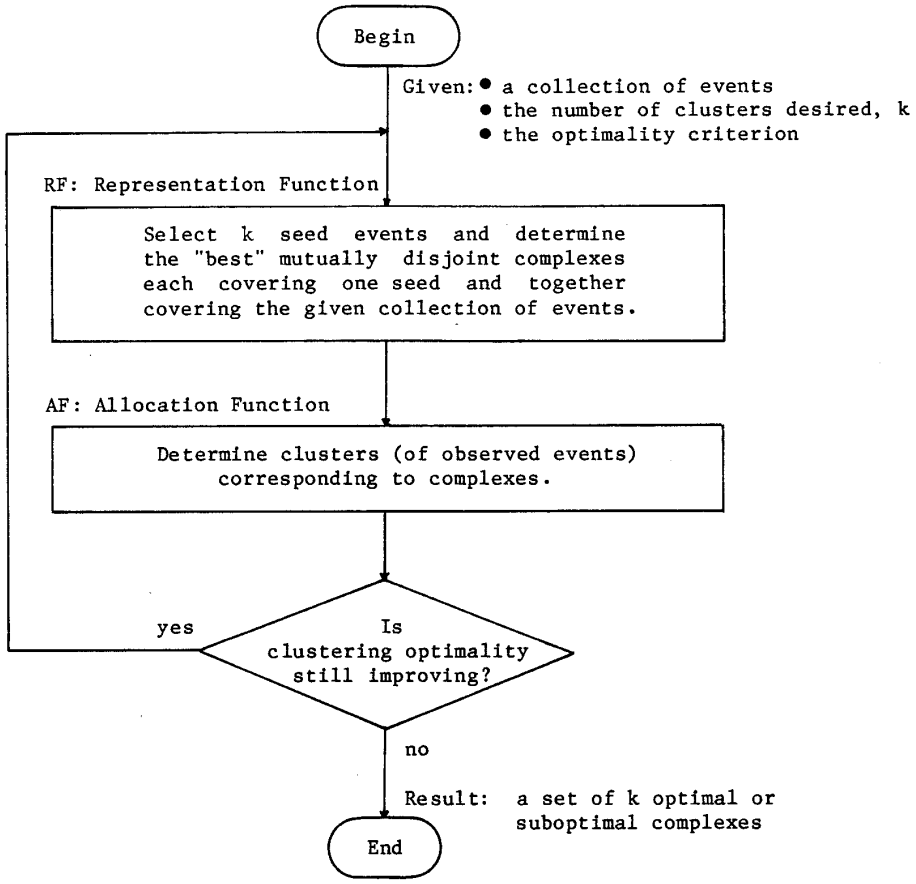


Figure 5. Simplified Flow of the Conjunctive Conceptual Clustering Algorithm

In the actual implementation, the program stores not only the best (locally optimal) k-clustering, but also a user-specified number of alternative k-clusterings closest to the best one from the viewpoint of the assumed clustering optimality criterion. Along with the k-clusterings (sets of k clusters) the program provides descriptions of individual clusters (ℓ-complexes) and their scores on the measures used in the optimality criterion (section IV). A detailed explanation of the algorithm is given in Michalski and Stepp [26,27]. A proof that every object collection can be partitioned into an arbitrary number of conjunctive concepts is in Michalski [25]. In Diday et al. [12] a clustering algorithm is described which uses an adaptive distance measure for creating clusters represented by a collection of complexes.

VI. An Example Problem

The simple example problem described below is used to illustrate some of the differences between conjunctive conceptual clustering and methods of numerical taxonomy.

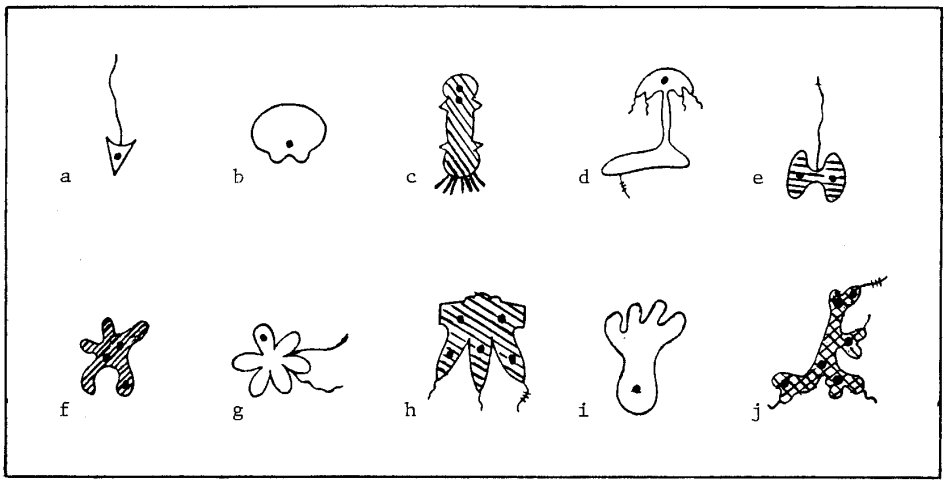


Figure 6. Microorganisms

The method-independent components of the problem (see section IIB) are:

1. The set of objects to be clustered: "microorganisms" shown in Figure 6.

2,3. The variables selected for describing microorganisms and their domains are:

Body parts
- 1 part
- 2 parts
- many parts

Body spots
- one spot
- many spots

Texture
- blank
- striped
- crosshatched

Tail type
- none
- single
- multiple

Figure 7 shows the descriptions of the microorganisms in terms of these variables.

4. The principle for grouping objects into clusters:
For numerical taxonomy: 18 different techniques are used, being combinations of three different similarity measures (product-moment correlation, simple matching coefficients, reciprocal Euclidean distance), three data transformations (none, normalizing variables into unit intervals, z-scores), and two clustering schemes (average linkage, and weighted average linkage).
For conjunctive conceptual clustering: the technique described in sections IV and V.

5. The inter-class structure: the partition structure.

Two programs were applied to solve this problem:

1. NUMTAX, developed by Professor Selander at the University of Illinois, which implements the 18 techniques of numerical taxonomy mentioned above (described in Sokal and Sneath [35]),

2. PAF, which implements conjunctive conceptual clustering (Michalski and Stepp [26,27]).

Micro-organism	Body parts	Body spots	Texture	Tail type
a	1	one	blank	single
b	1	one	blank	none
c	1	many	striped	multiple
d	2	one	blank	multiple
e	2	many	striped	single
f	many	many	striped	none
g	many	one	blank	multiple
h	many	many	striped	multiple
i	many	one	blank	none
j	many	many	crosshatched	multiple

Figure 7. Descriptions of microorganisms

Results from NUMTAX

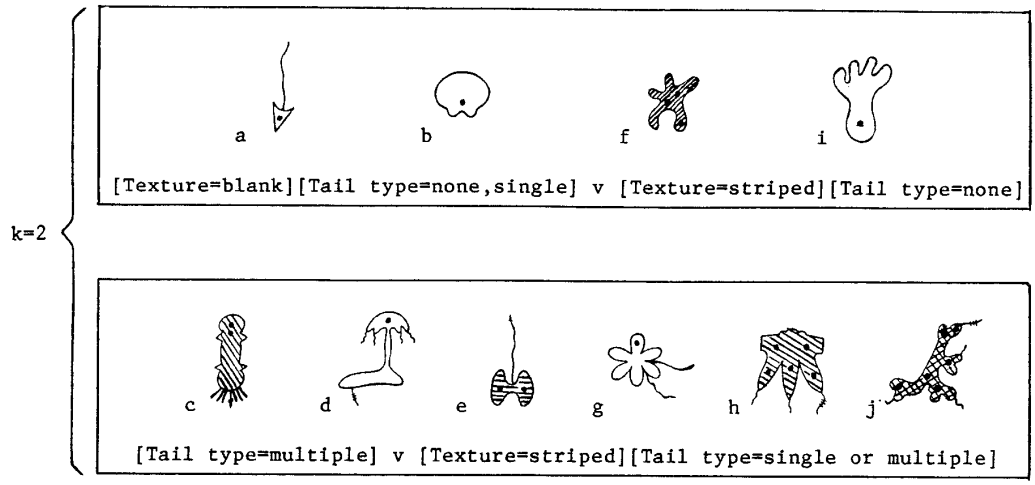
The numerical taxonomy program NUMTAX organizes the events into a hierarchy (a dendrogram) of clusters reflecting the numerical distances between consecutively larger clusters. The top level of the hierarchy represents the complete collection of events.

Because dendrograms are constructed bottom-up, the entire dendrogram must always be generated. After this is done, the dendrogram may be cut apart at some level to produce clusters. In our experiment, 18 different dendrograms were obtained, one from each technique applied (as described above). Figures 8 and 9 show two- and three-cluster solutions obtained from two typical dendrograms. The first dendrogram (Figure 8) was generated using the reciprocal of Euclidean distance, non-transformed data, and average linkage; the second dendrogram (Figure 9) was generated using using simple matching scores, transformed (z-score) data, and average linkage.

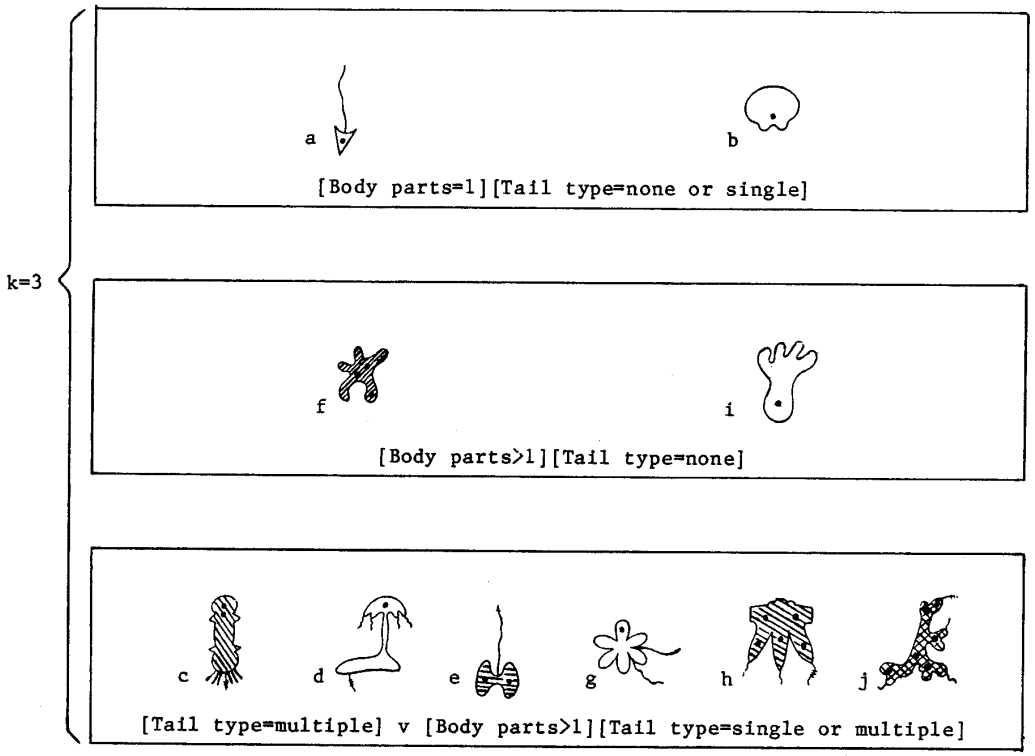
The clusters obtained from the dendrogram are not accompanied by any description. In order to determine descriptions of these clusters, they were presented to an inductive learning program, AQ11 (Michalski and Larson [24]), which formulated discriminant descriptions of each cluster (i.e., the shortest descriptions sufficient to discriminate between the clusters) in the form of a logical disjunction of VL_1 complexes. Descriptions listed in Figures 8 and 9 are results from the above program.

Results from conceptual clustering

Program PAF was run using the clustering optimality criterion: "maximize the essential dimensionality, then maximize the simplicity of cluster representations" (both with zero tolerance). The clusters and their descriptions obtained by PAF



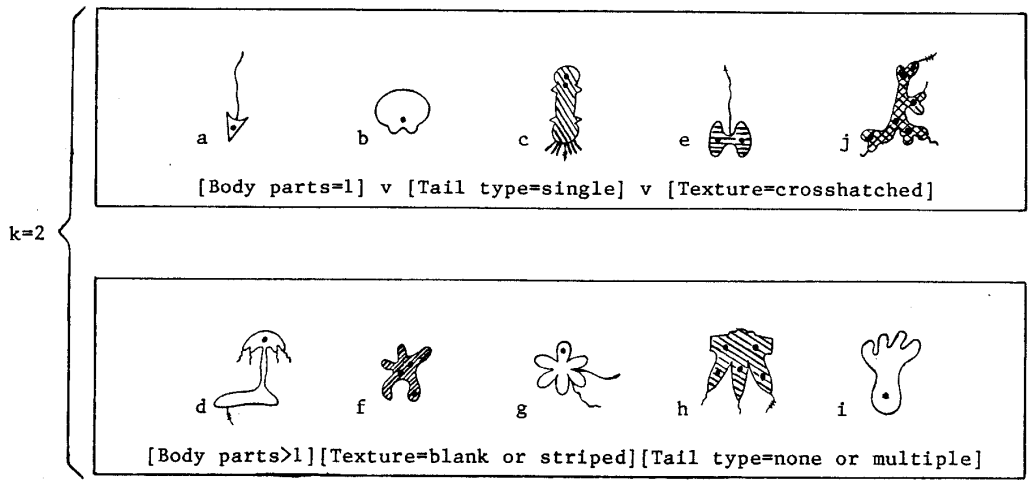
a.



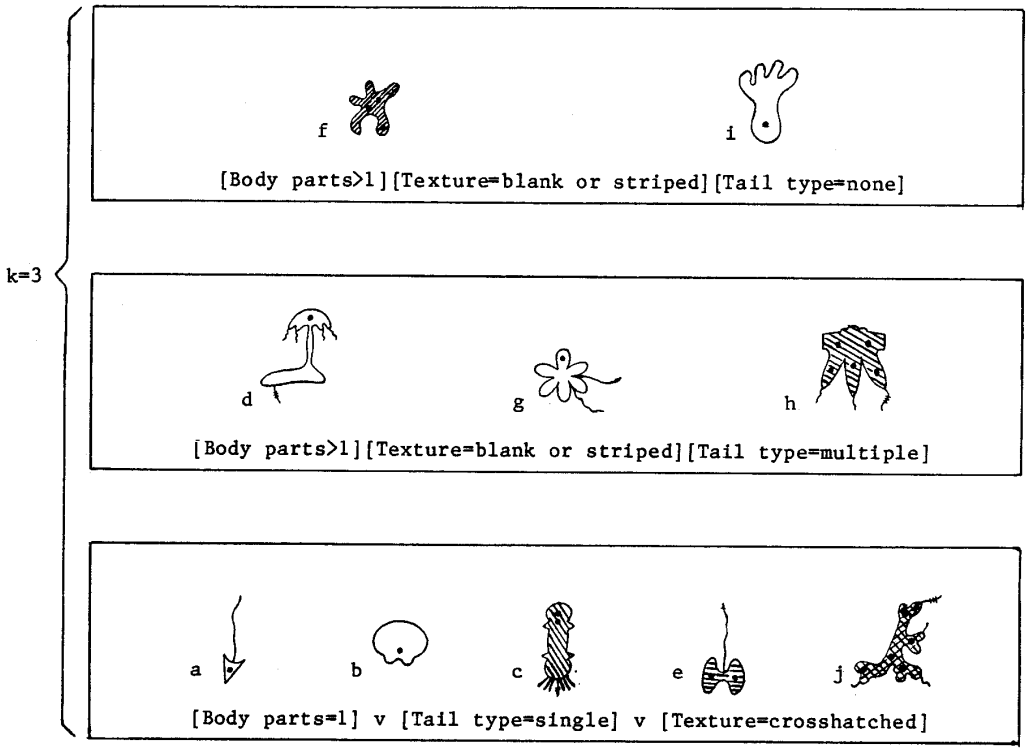
b.

(Cluster descriptions were produced by inductive program AQ11)

Figure 8. Clusters obtained by NUMTAX using average linkage, Euclidean distance and raw (non-transformed) data, for k=2 and k=3



a.



b.

(Cluster descriptions were produced by inductive program AQ11)

Figure 9. Clusters obtained by NUMTAX using average linkage, simple matching coefficients and transformed (z-score) data, for k=2 and k=3

with k=2 are shown in Figure 10a, and with k=3 in Figure 10b.

Discussion of results

An experiment with human subjects solving this problem indicated that people catagorized objects using the objects' most noticeable properties. Most common two-cluster solutions were:

(1) [Texture=blank] vs. [Texture≠blank], and
(2) [Body spots=one] vs. [Body spots=many]

and the most common three-cluster solution was:

(3) [Tail type=none] vs. [Tail type=single] vs. [Tail type=multiple]

When compared with the above solutions, the clusterings produced by NUMTAX seem rather arbitrary: the descriptions of single clusters (determined by program AQ11) involve disjunction in several cases, and are relatively complex. The descriptions produced by PAF, however, correspond well to human solutions. The program found that clusterings (1) and (2) are in fact identical. These human descriptions can be obtained directly from the descriptions generated by PAF by removing from PAF descriptions the conditions unnecessary for discriminating between the clusters. (In non-trivial problems this reduction is performed by applying inductive program AQ11). For k=3, PAF found, in addition to the solution shown in Figure 10b, the alternative solution:

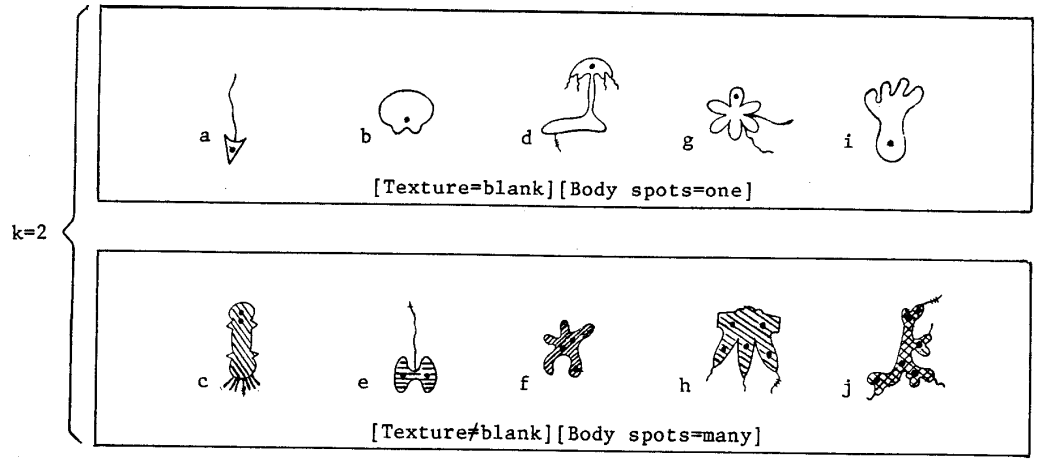
[Body parts=1][Texture=blank or striped]
[Body parts=2][Texture=blank or striped][Tail type=single or multiple]
[Body parts=many][Tail type=none or multiple]

Of the 18 dendrograms generated, only 4 (those involving either normalization or z-score data transformation, Euclidean distance, and either average or weighted average linkage) yielded a partitioning of data that matched the human solution (and PAF's solution, Figure 10a). Thus, in our experiments, numerical taxonomy methods produced clusters that in the majority of cases seemed to be rather inadequate from the viewpoint of human interpretation. This can be explained by noting that program NUMTAX is not equipped with any knowledge of human "concepts" (i.e., it does not know what types of solutions are preferred by humans), and therefore cannot knowingly produce clusters corresponding to such concepts.

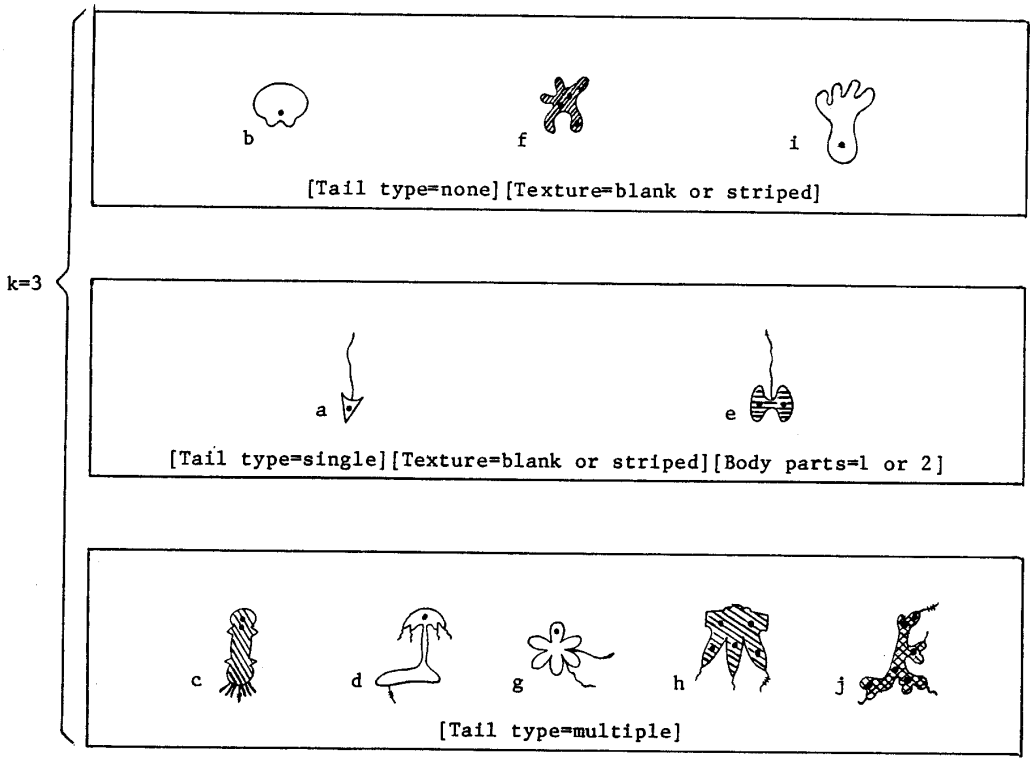
PAF has been tried on several practical clustering problems. One application was to cluster data describing 47 diseased soybean plants (each characterized by 34 many-valued variables). PAF accurately partitioned the diseased plants into four disease categories which were present in the sample, and described the clusters in terms of disease symptoms that agreed with the symptoms indicated by plant pathologists for these diseases.

VII. Conclusion

A method of conceptual clustering was discussed that produces clusters together with their descriptions in the form of conjunctive statements closely "fitting" the clusters. The important difference between this method and traditional clustering methods is that it does not use a similarity measure for forming clusters. Clusters are defined as groups of objects whose descriptions are disjoint logical products of relations on object attributes, optimizing a predefined criterion. Experiments performed so far have shown that the method produces clusters that tend to match solutions most satisfactory for people. Similar experiments with numerical taxonomy methods resulted in clusters that were less satisfactory in this regard.



a.



b.

Figure 10. Clusters and cluster descriptions obtained by PAF for k=2 and k=3, using as the optimality criterion: "maximize the essential dimensionality, then maximize the simplicity of cluster representations"

From the viewpoint of traditional clustering methods conceptual clustering can be interpreted as an approach that also uses a measure of object "similarity," but of a quite differend kind. This new kind of "similarity measure" takes into consideration not only the distance between objects (as in conventional clustering methods), but also their relationship to other objects and, most importantly, their relationship to some predetermined "concepts" (e.g., conjunctive concepts used in this paper).

The price of using such a more complex similarity measure is the significantly greater computational complexity of the method. For example, each dendrogram produced by NUMTAX required about 60 milliseconds of processor time on a CYBER 175, while clusterings produced by PAF required 1.5 to 4 seconds of processor time. (The above comparison is not totally appropriate because NUMTAX produces only clusters, while PAF produces both clusters and their descriptions.) The greater computational complexity is not necessarily a significant disadvantage of the method. If the results are indeed useful and practical, then the computational cost is of little relevance (especially now when the prices of computer technology are declining). Experience shows that researchers using presently available clustering techniques are most concerned not with the amount of computational time expended but with the difficulty of interpreting the results of the analysis. Another important characteristic and limitation of the implemented method is that it is oriented toward problems in which objects are described on nominal or ordinal scales (although it also handles variables measured on other scales after an appropriate quantization).

Concluding, the presented method of conjunctive conceptual clustering seems to add a new dimension to research in cluster analysis, and to have the potential to be a useful tool for researchers analyzing data.

VIII. Acknowledgements

The authors wish to thank Professor Richard Selander, University of Illinois Department of Genetics, for providing the numerical taxonomy program NUMTAX used in the comparative analysis of clustering methods. Partial support of this research was provided by the National Science Foundation grant No. MCS-79-06614, the University of Parix IX-Dauphine, and the Institut National de Recherche en Informatique et en Automatique (INRIA).

---

[1]Descriptions are disjoint if there are no events (observed or unobserved) that satisfy more than one description.

REFERENCES

[1] Anderberg, M. R., Cluster analysis for applications, (Academic Press, 1973).

[2] Backer, E., Cluster analysis formalized as a process of fuzzy identification based on fuzzy relations, Delft University of Technology, Department of Electrical Engineering, Report IT-78-15, October 1978.

[3] Ball, G. H., A clustering technique for summarizing variate data, Behavorial Science 12, No. 12, p. 153-155, 1967.

[4] Charles, C., Lechevallier, Y., Pattern recognition by a piecewise polynomial approximation with variable points, IRIA Laboria, Report No. 338, 1979.

[5] Chernoff, M., Metric consideration in cluster analysis, Proc. 6th Berkley Symposium on Math. Statistics and Probability, 1970.

[6] Coleman, G. B., Scene segmentation by clustering, University of Southern California, Image Processing Institute, report USCIPI, 1977.

[7] Cormark, R. M., A review of classification, Journal Royal Statictical Soc., series A, p. 134-321, 1971.

[8] Diday E., Govaert, G., Apprentissage et mesures de ressemblances adaptatives, Computer Oriented Learning Processes, Nato Advanced Study Institute, series E, No. 14, 1976.

[9] Diday, E. and Simon, J. C., Clustering analysis, Communication and Cybernetics 10, Springer-Verlag, Berlin, Heidelberg, New York, 1976.

[10] Diday, E. Problems of clustering and recent advances, 11th Congress of Statistics, Oslo, Norway, 1978.

[11] Diday E., et al, Optimization en classification automatique, INRIA, 1980.

[12] Diday, E., Govaert, G., Lechevallier, Y., Sidi, J., Clustering in pattern recognition, Fifth International Conference on Pattern Recognition, Miami Beach Florida, December 1-4, 1980.

[13] Dorofeyuk, A., Pattern recognition and machine learning without reward, Problems of Technical Cybernetics, Nauka, 1966.

[14] Do-Tu, H., Installe, M., A fast algorithm procedure based on ISODATA algorithm with application to remote sensing, 4th Int. Conf. on Pattern Recognition, Kyoto, Japan, p. 326, 1978.

[15] Fukada, Y., Spatial clustering procedures for region analysis, 4th Int. Conf. on Pattern Recognition, Kyoto, Japan, p. 329, 1978.

[16] Haralick, R. M., Shapiro, L., Decomposition of polygonal shapes by clustering, Proc. IEEE Conf. On Pattern Recognition and Image Processing, Troy, New York (USA), p. 183, 1977.

[17] Kasvand, T., Scene Segmentations and segment clustering experiments, 4th Int. Conf. on Pattern Recognition, Kyoto Japan, p. 426, 1978.

[18] Lechevallier, Y., Classification automatique sous contrainte d'ordre total, Rapport de recherche IRIA-Laboria No. 200, 1976.

[19] Lowitz, G. E., Compression des donnees par reconnaissance des formes et clustering, Congres AFCET-IRIA, p. 699, 1978.

[20] MacQueen, J., Some methods for classification analysis of multivariate observations, 5th Berkley Symposium on Mathematics, Statistics, and Probabilities, vol. 1, No. 281, 1967.

[21] Michalski, R. S., VARIABLE-VALUED LOGIC: System VL_1, Proceedings of the 1974 Intern. Symp. on Multiple-Valued Logic, West Virginia University, Morgantown, West Virginia, May 29-31, 1974.

[22] Michalski, R. S., Synthesis of optimal and quasi-optimal variable-valued logic formulas, Proceedings of the 1975 Intern. Symp. on Multiple-Valued Logic, Bloomington, Indiana, May 13-16, 1975.

[23] Michalski, R. S., Variable-valued logic and its applications to pattern recognition and machine learning, in: Rine, D. (ed.), Multiple-Valued Logic and Computer Science, North-Holland, 1975.

[24] Michalski, R. S. and Larson, J.B., Selection of most representative training examples and incremental generation of VL_1 hypotheses: the underlying methodology and the description of programs ESEL and AQ11, Report No. 867, Department of Computer Science, University of Illinois, Urbana, Illinois, 1978.

[25] Michalski, R. S., KNOWLEDGE ACQUISITION THROUGH CONCEPTUAL CLUSTERING: A theoretical framework and an algorithm for partitioning data into conjunctive concepts, A Special Issue on Knowledge Acquisition And Induction, Policy Analysis and Information Systems, No. 3, 1980.

[26] Michalski, R. S., Stepp, R., Revealing conceptual structure in data by inductive inference, in Machine Intelligence 10, eds. J. E. Hayes-Michie, D. Michie, and Y.-H. Pao, Chichister: Ellis Horwood, New York: Halsted Press (John Wiley), 1981.

[27] Michalski, R. S., Stepp, R., An application of AI techniques to structuring objects into an optimal conceptual hierarchy, Proceedings of the 7th International Joint Conference on Artificial Intelligence, Vancouver, Canada, August 24-28, 1981.

[28] Meisel, W., Computer oriented approaches to pattern recognition (Academic Press, 1972).

[29] Nilsson, Nils, T., Principles of artificial intelligence, Tioga Publishing Company, 1980.

[30] Nubuyaki, O., Discriminant and least squares threshold selection, 4th Int. Conf. on Pattern Recognition, Kyoto, Japan, p. 592, 1978.

[31] Pratt, W. K., Quantitive approaches to image feature extraction and segmentation, Congres AFCET-IRIA, p. 897, 1978.

[32] Roche, C., Rebuffet, M., Methodes de classification d'imagerie multi-spectrale, Congres AFCET-IRIA, p. 715, 1978.

[33] Rohlf, F. J., Adaptative hierarchical clustering schemas, Systematic Zoology, vol. 18, p. 58-82, 1970.

[34] Rosenfeld, A., Some recent developments in Texture Analysis, Proc. of Pattern Recognition and Image Processing, Chicago, 1979.

[35] Sokal, R. R., Sneath, P. H., Principles of numerical taxonomy (W. H. Freeman, 1963).

[36] Swain, P. H., Image data analysis in remote sensing, Digital Image Processing and Analysis, ed. Haralick and Simon, Leyder, Nordhoff, 1979.

[37] Taleng, Codage optimal adaptatif, Optimisation en classification automatique, Diday et al. Ed. INRIA, 1980.

[38] Yokoya, N., Kitahashy, T., Tanaka, K., Asano, T., Image segmentation schema based on a concept of relative similarity, 4th Int. Conf. on Pattern Recognition, Kyoto, Japan, p. 645, 1978.

[39] Zagoruiko, N. G., Methods of pattern recognition and their applications, Sovietskoie Radio, 1972 (in Russian).

[40] Zagoruiko, N. G., Lbov, G. S., Algorithms of pattern recognition in a package of applied programs, 4th Int. Conf. on Pattern Recognition, Kyoto, Japan, p. 1100, 1978.