KYBERNETIKA — VOLUME 17 (1981), NUMBER 1

LOGICALLY ORIENTATED CLUSTER ANALYSIS

ARNOŠT VESELÝ

This paper deals with a new logically orientated method of cluster analysis. Entities are grouped into clusters in such a way that among properties describing these entities strong logical relations are valid in the range of each cluster. The definition of the set of clusters is given and some properties of this set of clusters are considered. At the end an algorithm which yields an approximation of this set of clusters is described.

1. INTRODUCTION

The general aim of cluster analysis is to create sets of entities, so called clusters, in which each entity resembles all others of the same cluster. Suppose that each entity can be described by n binary parameters (properties). The fact that entity x has (resp. has not) k-th property we shall describe by the statement $V_k(x)$ (resp. $\overline{V}_k(x)$). In this way each entity can be characterized by a vector $\mathbf{v}(x) = (v_1(x), \ldots, v_k(x), \ldots, v_n(x))$, where $v_k(x) = 1$ iff statement $V_k(x)$ is valid or $v_k(x) = 0$ iff the opposite statement $\overline{V}_k(x)$ is valid.

The similarity of vectors $\mathbf{v}(x), \mathbf{v}(y)$ is usually measured (see [1]) by means of difference

$$|\mathbf{v}(x) - \mathbf{v}(y)| = \sum_{i=1}^{n} |v_i(x) - v_i(y)| . \tag{1.1}$$

If entities $x \in X$ are classified into classes $X_1, \ldots, X_m$ by means of some classifying algorithm, the most often aim is to classify the set X in such a way that

$$G(X_1, \ldots, X_m) = \sum_{j=1}^{m} \sum_{x,y \in X_j} |\mathbf{v}(x) - \mathbf{v}(y)| . \tag{1.2}$$

achieves its minimum (see [1]). Regrettably, it is not possible to achieve the minimum of (1.2) in the most real situations. Then it is necessary to use its attainable approximation.

Suppose $x \in X$ be entities examining in some experimental science. For example these entities might be patients and $V_k(x)$, $k = 1, 2, \ldots, n$ their symptoms. Classification of entities $x \in X$ into classes $X_1, \ldots, X_m$ is an important part of all experimental sciences. Some classification must be performed at the very beginning of its development and later this classification is modified and made more sophisticated. In medicine pacients are classified for example according to their diagnoses, syndroms etc.

Now we turn our attention to the definition of classes in experimental sciences. Classes are not usually defined by stating which logical formula consisting of properties $V_{i_1}, \ldots, V_{i_p}$ must be valid if an entity x is to be in certain class X_j. It means they are not defined by an expression of the following type

$$x \in X_j \equiv \Phi(V_{i_1}(x), \ldots, V_{i_p}(x)), \tag{1.3}$$

where $\Phi(V_{i_1}(x), \ldots, V_{i_p}(x))$ is a logical formula consisting of properties $V_{i_1}, \ldots, V_{i_p}$. R. Carnap [2] noticed that classes are often described by formulae of the following two types

$$\Phi_1 \supset (\Phi_2 \supset x \in X_j) \tag{1.4}$$

and

$$\Phi_3 \supset (\Phi_4 \supset (\overline{x \in X_j})), \tag{1.5}$$

where the formulae Φ_1, Φ_2, Φ_3, Φ_4 are arbitrary formulae consisting of properties $V_1, \ldots, V_n$. Following [2], formulae of the type (1.4) and (1.5) will be called reduction sentences of sentence $x \in X_j$. Similarly the pair of formulae of the type (1.4) and (1.5) will be called the reduction pair of sentences $x \in X_j$. Most usual there are formulae of the type $x \in X_j \supset \Phi$, which are equivalent to the formulae of the type

$$\bar{\Phi} \supset (\overline{x \in X_j}). \tag{1.6}$$

It is possible to consider them as special types of the reduction sentences (1.5), because formula (1.6) is equivalent to the formula $\Psi \supset (\Phi \supset (\overline{x \in X_j}))$ where Ψ is arbitrary logically valid formula.

From this point of view it is possible to say that in experimental sciences entities are classified in a such manner that for their properties within certain class a number of logical expressions are valid. Or one can say that among properties of entities of the same class there exists dependence which can be specified by logical expressions valid whithin this class.

In usual methods of cluster analysis the entities, which mostly resemble each other, are put into the same cluster. Resemblence in some sense is the most usual criterion of grouping entities together. In the following we shall state the requirement to group entities into clusters in such a way that among the properties describing entities strong relations are valid within each cluster.

2. DEFINITION OF CLUSTERS

Properties of elements $x_1, x_2, \ldots$ of a set X will be described by formulae of propositional calculus. If all elements $x \in X$ have a property V_i we shall say that the atomic formula V_i is valid. (We denote a property and a corresponding formula by the same symbol to simplify our notation; if a property or a formula is concerned will be obvious from the context.) Formulae of propositional calculus consisting of atomic formulae V_i and logical connectives $^-$, $\cdot$, $\vee$, $\supset$, $\equiv$ will represent assertions about properties of elements of the set X. For example, if the formula $(V_1 \vee V_2) \cdot \overline{V}_3$ is valid then all elements $x \in X$ have the property V_1 or V_2 and that no element $x \in X$ has the property V_3. Formulae will be denoted by capital letters of Greek alphabet Φ, Ψ, ...

From two properties V_1, V_2 one can create any number of formulae. For example formulae $V_1 \vee V_2$, $(V_1 \vee V_2) \vee V_2$, ... Some of them are equivalent. Formulae Φ_1 and Φ_2 are equivalent iff formula $\Phi_1 \equiv \Phi_2$ is tautology, i.e. iff $\vdash \Phi_1 \equiv \Phi_2$. For example formulae $\overline{V}_1 \vee \overline{V}_2$ and $\overline{V_1 \cdot V_2}$ are equivalent. Also formulae $V_1 \vee V_2$ and $(V_1 \vee V_2) \vee V_2$ are equivalent etc.

Lemma 2.1. Every formula consisting of properties V_1 and V_2 only is equivalent to one of the following formulae:

A.1. $V_1 \cdot V_2$,
A.2. $V_1 \cdot \overline{V}_2$,
A.3. $\overline{V}_1 \cdot V_2$,
A.4. $\overline{V}_1 \cdot \overline{V}_2$,
B.1. V_1,
B.2. $\overline{V}_1$,
B.3. V_2,
B.4. $\overline{V}_2$,
B.5. $V_1 \cdot V_2 \vee \overline{V}_1 \cdot \overline{V}_2$,
B.6. $V_1 \cdot \overline{V}_2 \vee \overline{V}_1 \cdot V_2$,
C.1. $V_1 \vee V_2$,
C.2. $V_1 \vee \overline{V}_2$,
C.3. $\overline{V}_1 \vee V_2$,
C.4. $\overline{V}_1 \vee \overline{V}_2$,
D.1. $V_1 \cdot \overline{V}_1$,
D.2. $V_1 \vee \overline{V}_1$.

Proof. Every formula must be equivalent to the one of 16 formulae in disjunctive normal form, which can be formed by connection of the rows of truth-value table in different way. After their minimalisation we obtain the above mentioned formulae. □

Formulae of group D are not interesting. Formula D.1. is not fulfilled in any class of entities and formula D.2. is fulfilled in all non-empty classes.

Lemma 2.2. For every formula Ψ of group B there exists formula Φ of group A such that $\vdash \Phi \supset \Psi$. Likewise for every formula Ψ of group B there exists formula Θ of group C such that $\vdash \Psi \supset \Theta$.

Proof. For every formula of group B we shall find the corresponding formula of group A. For example for $\overline{V}_1$ it may be $\overline{V}_1 \cdot V_2$ since $\vdash \overline{V}_1 \cdot V_2 \supset \overline{V}_1$.

In the similar fashion for every formula of group C we shall find the corresponding formula of group B. For example, for $V_1 \vee V_2$ it may be formula $V_1 \,.\, \overline{V}_2 \vee \overline{V}_1 \,.\, V_2$ since $\vdash V_1 \,.\, \overline{V}_2 \vee \overline{V}_1 \,.\, V_2 \supset V_1 \vee V_2$. □

Lemma 2.3. For none formula Φ of group A there exists formula Ψ of group B such that $\vdash \Psi \supset \Phi$. Likewise for none formula Ψ of group B there exists formula Θ of group C such, that $\vdash \Theta \supset \Psi$.

Proof. This lemma could be proved by evaluating truth-values tables for all possibilities.

Let $V_1, \ldots, V_n$ be n properties. The number of different two element subsets $\{V_j, V_k\}$ of the set $\{V_1, \ldots, V_n\}$ is $(n/2)$. We shall assign to every subset $\{V_j, V_k\}$ in the range of class X_l the weight $w_l(V_j, V_k)$ in the following way:

1) $w_l(V_j, V_k) = \delta_1$ iff some formula of group A (consisting of properties V_j, V_k) is valid in X_l.

2) $w_l(V_j, V_k) = \delta_2$ iff some formula of group B is valid in X_l and no formula of group A is valid in X_l.

3) $w_l(V_j, V_k) = \delta_3$ iff some formula of group C is valid in X_l and no formula of group A or B is valid in X_l.

4) $w_l(V_j, V_k) = \delta_4$ iff formulae of group D.2. only are valid in X_l (i.e. formulae $V_i \vee \overline{V}_i$ and $V_j \vee \overline{V}_j$). □

Suppose $\delta_1 < \delta_2 < \delta_3 < \delta_4$ are positive real numbers. It follows from Lemma 2.1, 2.2 and 2.3 that to every $\{V_j, V_k\}$ one and only one of the numbers $\delta_1, \delta_2, \delta_3, \delta_4$ is in the range of X_l assigned.

Definition 2.1. Let X be a set of entities and $V_1, \ldots, V_n$ their properties. For every disjoint system $\{X_1, \ldots, X_m\}$ of $X(\bigcup_{i=1}^{m} X_i = X,\ X_i \cap X_j = \emptyset$ for all $i, j = 1, \ldots, m;$ $i \neq j)$ and for any real positive numbers $\delta_1 < \delta_2 < \delta_3 < \delta_4$ we shall define $F_{\delta_1,\ldots,\delta_4}(X_1, \ldots, X_m)$ in the following way:

$$F_{\delta_1,\ldots,\delta_4}(X_1, \ldots, X_m) = \sum_{i=1}^{m} \sum_{j,k=1;j\neq k}^{\binom{n}{2}} w_l(V_j, V_k)\,. \tag{2.1}$$

Set of m clusters $\{C_1, \ldots, C_m\}$ of the set X is defined as a such disjoint system of subsets of X for which (2.1) achieves its minimum value, i.e. for which

$$F_{\delta_1,\ldots,\delta_4}(C_1, \ldots, C_m) = \min_{\{X_1,\ldots,X_m\}\in X^m} F_{\delta_1,\ldots,\delta_4}(X_1, \ldots, X_m) \tag{2.2}$$

is valid.

Thus it follows by Definition 2.1 that set of clusters is that decomposition $\{X_1, \ldots, X_m\}$ of X, in which the strongest logical expressions consisting of properties V_j, V_k are valid.

Example 2.1. Let $X = \{x_1, x_2, x_3, x_4\}$ be a set and let every element of X be characterized by four properties V_1, V_2, V_3, V_4 according to the Table 2.1. The set

Table 2.1.

	V_1	V_2	V_3	V_4
x_1	0	0	0	0
x_2	0	1	0	0
x_3	1	1	0	0
x_4	1	0	1	1

Table 2.2.

X_1	X_2	$F(X_1, X_2)$	$G(X_1, X_2)$
$\{x_1\}$	$\{x_2, x_3, x_4\}$	21	8
$\{x_2\}$	$\{x_1, x_3, x_4\}$	23	8
$\{x_3\}$	$\{x_1, x_2, x_4\}$	21	8
$\{x_4\}$	$\{x_1, x_2, x_3\}$	18	4
$\{x_1, x_2\}$	$\{x_3, x_4\}$	21	4
$\{x_1, x_3\}$	$\{x_2, x_4\}$	23	6
$\{x_1, x_4\}$	$\{x_2, x_3\}$	21	4

Table 2.3.

	$\{x_1, x_2\}$	$\{x_1, x_3\}$	$\{x_1, x_4\}$	$\{x_2, x_3\}$	$\{x_2, x_4\}$
(V_1, V_2)	$\overline{V}_1$	$V_1 \cdot V_2 \vee \overline{V}_1 \cdot \overline{V}_2$	$\overline{V}_2$	V_2	$V_1 \cdot \overline{V}_2 \vee \overline{V}_1 \cdot V_2$
$w_t(V_1, V_2)$	2	2	2	2	2
(V_1, V_3)	$\overline{V}_1 \cdot \overline{V}_3$	$\overline{V}_3$	$V_1 \cdot V_3 \vee \overline{V}_1 \cdot \overline{V}_3$	$\overline{V}_3$	$V_1 \cdot V_3 \vee \overline{V}_1 \cdot \overline{V}_3$
$w_t(V_1, V_3)$	1	2	2	2	2
(V_1, V_4)	$\overline{V}_1 \cdot \overline{V}_4$	$\overline{V}_4$	$V_1 \cdot V_4 \vee \overline{V}_1 \cdot \overline{V}_4$	$\overline{V}_4$	$V_1 \cdot V_4 \vee \overline{V}_1 \cdot \overline{V}_4$
$w_t(V_1, V_4)$	1	2	2	2	2
(V_2, V_3)	$\overline{V}_3$	$\overline{V}_3$	$\overline{V}_2$	$V_2 \cdot \overline{V}_3$	$V_2 \cdot \overline{V}_3 \vee \overline{V}_2 \cdot V_3$
$w_t(V_2, V_3)$	2	2	2	1	2
(V_2, V_4)	$\overline{V}_4$	$\overline{V}_4$	$\overline{V}_2$	$V_2 \cdot \overline{V}_4$	$V_2 \cdot \overline{V}_4 \vee \overline{V}_2 \cdot V_4$
$w_t(V_2, V_4)$	2	2	2	1	2
(V_3, V_4)	$\overline{V}_3 \cdot \overline{V}_4$	$\overline{V}_3 \cdot \overline{V}_4$	$V_3 \cdot V_4 \vee \overline{V}_3 \cdot \overline{V}_4$	$\overline{V}_3 \cdot \overline{V}_4$	$V_3 \cdot V_4 \vee \overline{V}_3 \cdot \overline{V}_4$
$w_t(V_3, V_4)$	1	1	2	1	2
$\sum_{j=1,k=1,j\neq k}^{\binom{n}{2}} w_t(V_j, V_k)$	9	11	12	9	12

X is to be decomposed into two clusters according to the Definition 2.1 and under the assumption that $\delta_1 = 1$, $\delta_2 = 2$, $\delta_3 = 3$, $\delta_4 = 4$.

All possible disjoint systems $\{X_1, X_2\}$ of X, $(X_1 \cup X_2 = X, X_1 \cap X_2 = \emptyset)$ are in the first two columns of the Table 2.2. In the third column of this table there are values of $F(X_1, X_2)$. (In the following we shall omit $\delta_1, \ldots, \delta_4$ from the notation of $F_{\delta_1,\ldots,\delta_4}(X_1, \ldots, X_m)$ if $\delta_1 = 1$, $\delta_2 = 2$, $\delta_3 = 3$, $\delta_4 = 4$). The values of $F(X_1, X_2)$ are computed in the auxiliary Table 2.3. The values of $G(X_1, X_2)$ take place in the fourth column of the Table 2.2. $G(X_1, X_2)$ is defined by expression (1.2) and if $G(X_1, X_2)$ is used for decompositions of X, the usual set of clusters is obtained.

If usual definition of set of clusters is used, then all the following disjoint systems of X should be considered as sets of clusters: a) $\{x_1, x_2, x_3\}$, $\{x_4\}$; b) $\{x_1, x_4\}$, $\{x_2, x_3\}$; c) $\{x_1, x_2\}$, $\{x_3, x_4\}$. If the Definition 2.1 is used, then only decomposition $\{x_1, x_2, x_3\}$, $\{x_4\}$ is the set of clusters of X.

In the following we shall examine the problem of choice of values of $\delta_1, \ldots, \delta_4$. In the Example 2.1 we have chosen $\delta_1 = 1$, $\delta_2 = 2$, $\delta_3 = 3$, $\delta_4 = 4$. We shall show the result of this choice and we shall see that this choice is in some way natural.

Theorem 2.4. Let be $\delta_1 = 1$, $\delta_2 = 2$, $\delta_3 = 3$, $\delta_4 = 4$; $X_l = \{x_1, \ldots, x_m\}$ and let $\boldsymbol{M}(X_l)$ be a matrix

$$\begin{pmatrix} v_1(x_1), & v_2(x_1), & \ldots, & v_n(x_1) \\ \vdots & \vdots & & \vdots \\ v_1(x_m), & v_2(x_m), & \ldots, & v_n(x_m) \end{pmatrix}.$$

$\{x_3, x_4\}$	$\{x_1, x_2, x_3\}$	$\{x_1, x_2, x_4\}$	$\{x_2, x_3, x_4\}$	$\{x_1, x_3, x_4\}$
V_1	$\overline{V}_1 \vee V_2$	$\overline{V}_1 \vee \overline{V}_2$	$V_1 \vee V_2$	$V_1 \vee \overline{V}_2$
2	3	3	3	3
V_1	$\overline{V}_1$	$V_1 . V_3 \vee \overline{V}_1 . \overline{V}_3$	$V_1 \vee \overline{V}_3$	$V_1 \vee \overline{V}_3$
2	2	2	3	3
V_1	$\overline{V}_4$	$V_1 . V_4 \vee \overline{V}_1 . \overline{V}_4$	$V_1 \vee \overline{V}_4$	$V_1 \vee \overline{V}_4$
2	2	2	3	3
$V_2 . \overline{V}_3 \vee \overline{V}_2 . V_3$	$\overline{V}_3$	$\overline{V}_2 \vee \overline{V}_3$	$V_2 . \overline{V}_3 \vee \overline{V}_2 . V_3$	$\overline{V}_2 \vee \overline{V}_3$
2	2	3	2	3
$V_2 . \overline{V}_4 \vee \overline{V}_2 . V_4$	$\overline{V}_4$	$\overline{V}_2 \vee \overline{V}_4$	$V_2 . \overline{V}_4 \vee \overline{V}_2 . V_4$	$\overline{V}_2 \vee \overline{V}_4$
2	2	3	2	3
$V_3 . V_4 \vee \overline{V}_3 . \overline{V}_4$	$\overline{V}_3 . \overline{V}_4$	$V_3 . V_4 \vee \overline{V}_3 . \overline{V}_4$	$V_3 . V_4 \vee \overline{V}_3 . \overline{V}_4$	$V_3 . V_4 \vee \overline{V}_3 . \overline{V}_4$
2	1	2	2	2
12	12	15	15	17

Let $\boldsymbol{M}_{jk}(X_l)$ be a submatrix of $\boldsymbol{M}(X_l)$ consisting of its two columns j and k. Then the value of $w_l(V_j, V_k)$ is equal to the number of different rows of submatrix $\boldsymbol{M}_{jk}(X_l)$.

Proof. 1) Let be $w(V_j, V_k) = 1$. Then one of formulae of group A must be valid in classX_l. Let it be for example formula $V_j \,.\, V_k$. In this case, all rows of submatrix $\boldsymbol{M}_{jk}(X_l)$ must be vectors (1,1). Hence the value of $w_l(V_j, V_k)$ is equal to the number of different rows of submatrix $\boldsymbol{M}_{jk}(X_l)$. We should arrived at the same conclusion if we should take into consideration all other formulae of the group A, i.e. $V_1 \,.\, \overline{V}_2$, $\overline{V}_1 \,.\, V_2$ and $\overline{V}_1 \,.\, \overline{V}_2$.

2) For values 2, 3, 4 of $w_l(V_j, V_k)$ we should prove the assertion of this theorem in a similar fashion. □

An immediate consequence of the Theorem 2.4 is that $F(X_1, \ldots, X_m)$ equals to the sum

$$\sum_{l=1}^{m} \sum_{j,k=1;\, j \neq k}^{\binom{n}{2}} d(\boldsymbol{M}_{jk}(X_l)),$$

where $d(\boldsymbol{M}_{jk}(X_l))$ is the number of different rows of submatrix $\boldsymbol{M}_{jk}(X_l)$. Thus the set of m clusters $\{C_1, \ldots, C_m\}$ is that decomposion of X for which this sum is minimum.

3. CLUSTER ALGORITHM

In Sec. 2 the set of clusters was defined. Howewer, the question how to determine this set of clusters arises in practice. In the Example 2.1 of Sec. 2, at first, the values of $F(X_1, \ldots, X_m)$ for all possible decompositions of X were computed. Then according to (2.2) the decomposition of X, for which the value of $F(X_1, \ldots, X_m)$ was minimum, was taken as the set of clusters of X. In our example the set X consisted of 4 elements only. But in practical situations the set X might consist of hundreds of elements. If n denotes the number of elements of X, the number of all possible decompositions of X into two classes is 2^n. The number of all possible decompositions of X into more than two classes is much greater. This fact implies, that in the case $n \approx 100$, the set of clusters could not be determined in a such simple way as in Example 2.1. The generation of 2^n, $n \approx 100$ decompositions of set X would not be possible even if the most modern computer would be used.

It should be noted that the same difficulty arises when usual definition of set of clusters is used, i.e. if the set of clusters is defined as that decomposition of X for which (1.2) attains minimum. Even for this most usual definition of the set of clusters, it is not possible to determine the set of clusters, if $n \approx 100$. We must be content with an approximation attainable by means of a computer. Most often this approximation is the result of using ISODATA algorithm [3] or some of its modifications.

In the following we shall put down an algorithm which yields an approximation to the set of clusters as defined by Definition 2.1.

Cluster algorithm.

1. Let $x_1, x_2, \ldots, x_m$ be m arbitrary elements of X. Then put $X_1^0 = \{x_1\}, \ldots, X_m^0 = \{x_m\}$.

2. Let $X_1^r, \ldots, X_m^r$ be sets determined in the r-th step of the algorithm.

a) If the set $X - \bigcup_{i=1}^{m} X_i^r$ is non-empty, pick up an arbitrary element $x \in X - \bigcup_{i=1}^{m} X_i^r$ and compute

$$F^1(X_1^r \cup \{x\}, X_2^r, \ldots, X_m^r),$$
$$F^2(X_1^r, X_2^r \cup \{x\}, \ldots, X_m^r),$$
$$\vdots$$
$$F^m(X_1^r, X_2^r, \ldots, X_m^r \cup \{x\}).$$

Let s be the minimum natural number $1 \leqq s \leqq m$ for which

$$F^s = \min_{t \in \langle 1, m \rangle} F^t$$

Then put $X_i^{r+1} = X_i^r$ for all $i = 1, \ldots, m$; $i \neq s$ and $X_s^{r+1} = X_s^r \cup \{x\}$.

b) If the set $X - \bigcup_{i=1}^{m} X_i^r$ is empty, the algorithm stops and the set $\{X_1^r, \ldots, X_m^r\}$ is taken as an approximation to the set of clusters.

The algorithm described above can be easily modified in the following way. Item 2. will be executed for all $x \in (X - \bigcup_{i=1}^{m} X_i^r)$ and that x will be chosen, for which F^s is minimum. In the following Algorithm 1 denotes the primary algorithm and Algorithm 2 denotes its modification.

It is obvious, that if Algorithm 1 is used, the approximation of the set of clusters depends on the choice of $X_1^0, \ldots, X_m^0$ and on the sequence $x_1, \ldots, x_r, \ldots$, where $x_r \in X - \bigcup_{i=1}^{m} X_i^r$ is the element of X chosen in r-th step. If the Algorithm 2 is used, the approximation depends on the sequence $X_1^0, \ldots, X_m^0$ only. The realisation of Algorithm 2 needs much more computing than a realisation of Algorithm 1. Therefore we recommend to use the Algorithm 1 more times with different choice of initial sets $X_1^0, \ldots, X_m^0$ and with different sequences $x_1, \ldots, x_j, \ldots$. If no information about the set of clusters is given apriori, initial sets $X_1^0, \ldots, X_m^0$ must be chosen randomly. The elements x_j of a sequence $x_1, \ldots, x_j, \ldots$ are to be chosen randomly too.

To illustrate a behaviour of the Algorithm 1, we shall consider the Example 2.1. Initial sets let be $X_1^0 = \{x_1\}$ and $X_2^0 = \{x_2\}$. A sequence of elements, which are chosen in the step 2 of the algorithm, let be x_2, x_3. The process of formation of clusters is illustrated in Fig. 3.1. At every vertex sets X_1^r, X_2^r are put down and inside a ring, which denotes a vertex, a corresponding value of $F(X_1^r, X_2^r)$ is written. Between two verteces is put down that element x, which is used in the second step of the algorithm

for a formation of X_i^{r+1} of X_i^r. In Fig. 3.2 a formation of a set of clusters is illustrated for the same choice of initial sets X_1^0, X_2^0, but for a different sequence of elements chosen in the step 2 of the algorithm. In this case the sequence is x_3, x_2. From these two figures it is obvious that in the both cases the approximation of set of clusters, which is the result of the application of the Algorithm 1, is $\{C_1, C_2\}$ i.e. the set of clusters itself.

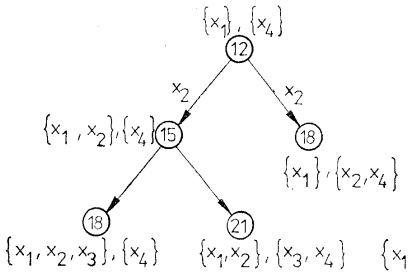

Fig. 3.1.

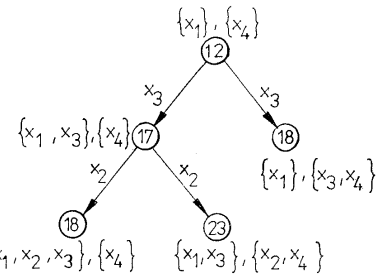

Fig. 3.2.

In the last example initial sets X_1^0 X_2^0 were chosen in a such way that $X_1^0 \subset C_1$ and $X_2^0 \subset C_2$. Suppose now $X_1^0 = \{x_1\}$ and $X_2^0 = \{x_2\}$. Then $X_1^0 \subset C_1$ and $X_2^0 \subset C_2$ cannot be both valid. The behavior of the algorithm in this case is illustrated in Fig. 3.3 for the sequence x_3, x_4 and in Fig. 3.4 for the sequence x_4, x_3. We know from Sec. 2 that the set $\{\{x_1, x_2, x_3\}, \{x_4\}\}$ is the only set of clusters of the set X. But the elements x_1, x_2 were put in the first step of the algorithm into different initial sets. Taking into consideration that during a run of the algorithm every element x once put into X_i^r remains there, the elements x_1 and x_2 must remain in different sets.

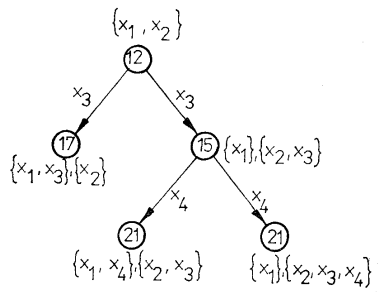

Fig. 3.3.

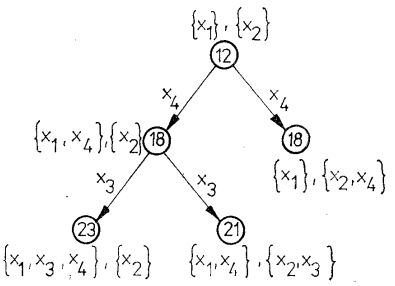

Fig. 3.4.

Therefore decomposition of X, which is a result of the algorithm cannot be the set of clusters $\{C_1, C_2\}$, but only its approximation. In effect, for both sequences x_3, x_4 and x_4, x_3 the resulting decomposition of X is the same. It is decomposition $\{\{x_1, x_4\}, \{x_2, x_3\}\}$.

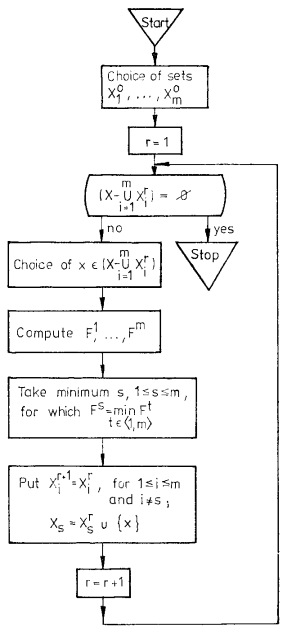

Fig. 3.5.

Notice that if the set of clusters is defined in the usual way, i.e. according to (1.2), then the decomposition $\{\{x_1, x_4\}, \{x_2, x_3\}\}$ is the set of clusters of X too. (See Table 2.2).

At the end we shall present flow chart of the Algorithm 1 (see Fig. 3.5). Using this flow chart a program for computer realising the Algorithm 1 could be work out.

(Received June 6, 1977.)

REFERENCES

[1] R. M. Cormack: A review of classification. J. Royal Statist. Soc. *134* (1971), 321—367.
[2] R. Carnap: Testability and meaning. Philosophy of Sciences (1936), 3, 419—471.
[3] G. H. Ball, D. J. Hall: ISODATA a Novel Method of Data Analysis and Pattern Classification. Tech. Rept., Stanford Research Inst., Menlo Park, Calif. 1965.

Ing. Arnošt Veselý, Oblastní výpočetní centrum VŠ (Regional University Computing Centre), Zikova 4, 166 29 Praha 6. Czechoslovakia.