Chapter 5
Clusterization and Recognition

1 SELF-ORGANIZATION MODELING AND CLUSTERING

The inductive approach shows that the most accurate predictive models can be obtained in the domain of nonphysical models that do not possess full complexity. This corresponds to Shannon's second limit theorem of the general communication theory. The principle of self-organization is built up based on the Gödel's incompleteness theorem. The term "self-organization modeling" is understood as a sorting of many candidates or partial models by the set of external criteria with the aim of finding a model with an optimal structure.

A "fuzzy" object is an object with parameters that change slowly with time. Let us denote N as a number of data points and m as a number of variables. For $N < m$, the sample is called short and the object "fuzzy" (under-determined). The greater the ratio m/N, the "fuzzier" the object.

By describing the relationships, clustering is considered a model of an object in a "fuzzy" language. Sorting of clusters with the aim of finding an optimal cluster is called "self-organization clustering." Although self-organization clustering has not yet been developed in detail, it has adapted the main principles and practical procedures from the theory of "self-organization modeling." This chapter presents the recent developments of self-organization clustering and nonparametric forecasting and explains how the principles of self-organization theory are applicable for identifying the structure of the most accurate and unbiased clusterizations.

Analogy with Shannon's approach

Structural identification by self-organization modeling is directed not only toward obtaining a physical model, but also toward obtaining a better, and not overly complicated, prediction model. The theoretical basis of this statement is taken from the communication theory by Shannon's second-limit theorem for transmission channels with noise. The optimal complexity of clusterizations is required as the optimal frequency passband in a communication system. Complexity must decrease as the variance of noise increases. The complexity of the models to be evaluated is often measured by the number of parameters and the order of the equation. The complexity of clusterization is usually measured by the number of clusters and attributes. The complexity of a model or clusterization is determined by the magnitude of the minimum-bias of the criterion as minimum of the Shannon-bias. The greater the bias, the simpler the object of investigation. The measurement of bias represents the difference of the abscissa of the characteristic point of the physical model. Bias is measured for different models of varying complexities. However, without Shannon's approach, it would be incomprehensible why one cannot find a physical model for noisy data and why a physical model is not suitable for predictions. This is analogous to the noise immunity of the criteria for template sorting in cluster analysis.

Godel and non-Godel types of systems

The inductive approach is fundamentally a different approach. It has a completely opposite assertion to the deductive opinion of "the more complex the model, the more accurate it is" with regard to the existence of a unique model with a structure of optimal complexity. It is possible to find an optimal model for identification and prediction only by using the external criteria.

The concept of "external criteria" is connected with the Gödel's incompleteness theorem. This means that the Godel type systems use a criterion realizing the support of the system on an external medium, which is like an external controller in a feedback control system. There is no such controller in the non-Godel type systems. Usually, the controller is replaced by a differential element for comparison of two quantities without any explicit reference to the external medium.

Let us recall some of the basic propositions of these theories of modeling. In case of ideal data (without noise), both approaches produce the same choice of optimal models or clustering with the same optimal set of features. In case of noisy data, the advantage with Godel's approach is that although the method is robust compared to the non-Godel type, it captures the optimal robust model or clustering with its basic features. It conveys to the modeler that it is simpler to follow traditional approaches without taking any complicated paths with inductive approaches. However, an obvious affirmative solution to this question, in which the training data sample does not participate, must be sought among external criteria.

One important feasibility of such a criterion that possesses the properties of an external controller is the partitioning of data sample into two subsets A and B by the subsequent comparison of the modeling or clustering results obtained for each of them. Various examples of constructing the criteria differ according to the initial requirement and in the degree of fuzziness of the mathematical language.

Division of data as per dipoles

In self-organization modeling, usually the data points with a larger variance of the output quantity are taken into the training set A and the points with a smaller variance are taken into the testing set B. Such a division is not applicable in self-organization clustering because "local clusters" of points for the subsamples are destroyed. The "dipoles" of the data sample as point separations allow us to find $(N/2— 1)$ pairs of points nearest to one another, where N denotes the total number of points in the sample. Figure 5.1 depicts six "dipoles" whose vertices are used to form the sets A and B, as well as C and D. The points located closer to the observation point / are taken into the set A, while those closer to the observation point // are taken into the set B. The other vertices of the dipoles respectively form the sets C and D. This is also demonstrated in one of the examples given in this chapter.

Clusterization using internal and external criteria

Cluster analysis is usually viewed as a theory of pattern recognition "without teacher"; i.e., without indication of a target function. The result of the process is called clusterization. We know that the theory of clustering is not a new one. One can find a number of clustering algorithms existing in pattern-recognition literature that allow clusterization to be obtained;

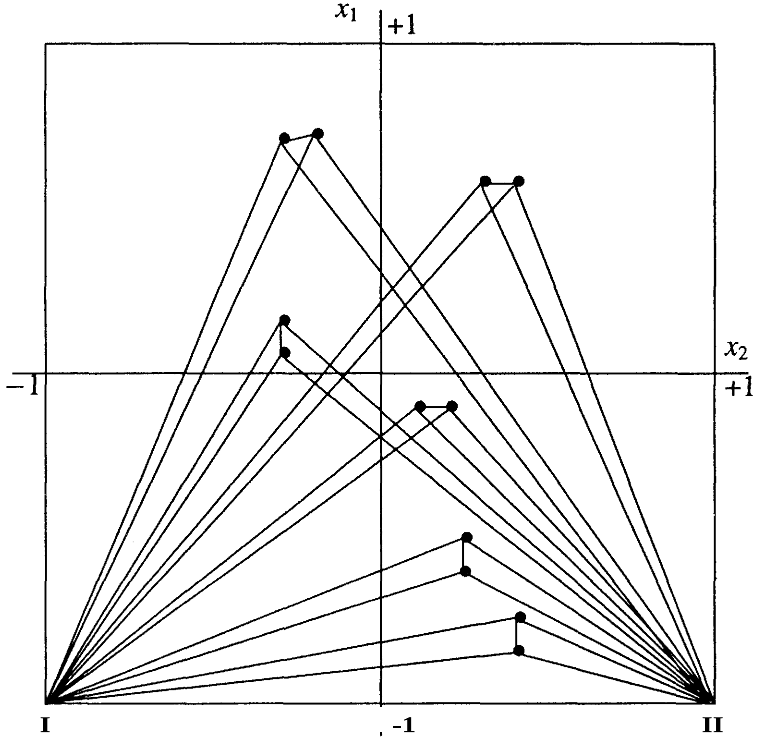


Figure 5.1. Partitioning of data sets A, B from observation I and C, D from observation II

namely, to divide a given set of objects represented by data points in a multi-dimensional space of attributes into a given number of compact groups or clusters. Most of the traditional algorithms are used in the formation of clusters and in the determination of their optimal number by using a single internal criterion having a meaning related to its accuracy or information. With a single criterion, we obtain "the more clusters—the more accurate the clusterization." It is needed for specifyng either a threshold or some constraints when the choice of the number of clusters is made.

Here it describes algorithms for objective computer clusterization (OCC) in which clusters are formed according to an internal, minimum-distance criterion. Their optimal number and the composition of attributes are determined by an external, minimum-bias criterion called a consistency or non-contradictory criterion. Any criterion is said to be external when it does not require specification of subjective thresholds or constraints. The criteria regularity (called precision or accuracy here), consistency, balance-of-variables, and so on, serve as examples of external criteria. Internal criteria are those that do not form the minimum, and therefore exclude the possibility of determining a unique model or clusterization in optimal complexity corresponding to global minimum.

Explicit and implicit templates

The main difference between self-organization modeling and self-organization clustering is the degree of detail of the mathematical language. In clustering analysis, one uses the language of cluster relationships for representing the symptoms and the distance measurements as objective functions instead of equations. The synthesis of models in the implicit form $f(x) = 0$ corresponds to the procedure of unsupervised learning (without teacher, in the literature it is also notified as competitive learning) and in the explicit form $y = f(x)$ it corresponds to the procedure of supervised learning (with teacher).

The objective system analysis (OSA) algorithm usually chooses a system that contains three to five functions which are clearly insufficient for describing large scale systems. Such "modesty" of the OSA algorithm is only superficial. Indeed, a small system of equations is basic, but the algorithm identifies many other systems which embrace all the necessary variables using the minimum-bias criterion. The final best system of equations is chosen by experts or by further sorting of the best ones. What one really has to sort in the inductive approach is not models, plans, or clusterings, but their explicit or implicit templates (Figure 5.2). This helps in the attainment of unimodality of the "criterion-template complexity." If the unimodality is ensured, then the characteristics look as they do in Figure 5.3 for different noise levels. The figures demonstrate the results of sorting of explicit and implicit templates; i.e., in single and system models, correspondingly. These are obtained by computational experiments that use inductive algorithms with regularity and consistent criteria. "Locus of the minima" represents the path across the minimum values achieved at each noise level.

Self-organization of clusterization systems

The types of problems we discuss here—one is the sorting of partial models and other is sorting of clusters—can be dealt with with some care and modeling experience. Figure 5.3b shows the curves that are characteristic for objective systems analysis. Here the model is represented not by a single equation, but by a system of equations, and one can see a gradual widening of the boundaries of the modeling region. There is a region which is optimal with respect to the criterion. The problem of convolution of the partial criteria of individual equations are encountered into a single system criterion.

The theory behind obtaining the system of equations also applies to clusterization in the form of partial clusterization systems that differ from one another in the set of attributes and output target functions. For example, in certain properties of the object, two independent autonomous clusterizations of the form

$$< y_1 > \leftrightarrow < x_{11}x_{12}x_{13}\cdots x_{1m} >, \quad < y_2 > \leftrightarrow < x_{21}x_{22}x_{23}\cdots x_{2m} >$$

have to be replaced by a system of two clusterizations being jointly considered

$$< y_1 > \leftrightarrow < x_{11}x_{12}x_{13}\cdots x_{1m}y_2 >, \quad < y_2 > \leftrightarrow < x_{21}x_{22}x_{23}\cdots x_{2m}y_1 >,$$

where y_1 and y_2 are the output components corresponding to certain properties of the object and x_{ij}, $(i = 1, 2$ and $j = 1, 2, \cdots, m)$ first denote two data points corresponding to the m input attributes.

This is analogous to the operation of going from explicit to implicit templates. The optimal number of partial clusterizations forming the system is determined objectively according to the attainable depth of the minimum of the criteria as achieved in the OSA algorithm.

Figure 5.4 illustrates the results of self-organization in sorting of clusterings by showing a special shape of curve using two criteria: consistency and regularity. The objective based self-organization algorithms are oriented toward the search for those clusterizations that are unique and optimal for each noise level, although the overall consistency criterion leads to zero as the noise variance is reduced. It is helpful to have some noise within the limits in the data; however, the greater the inaccuracy of the data, the simpler the optimal clusterization.

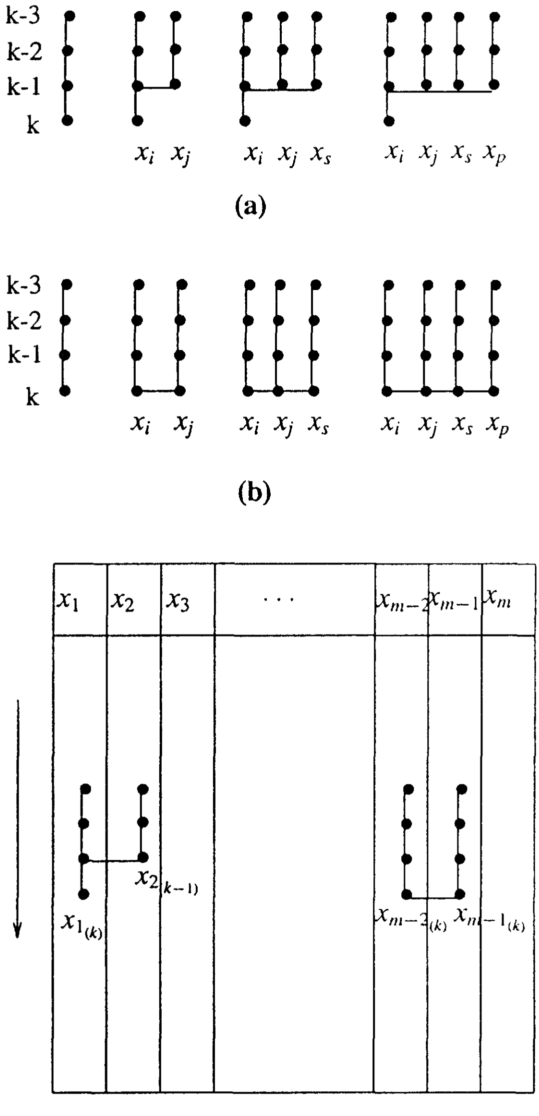


Figure 5.2. Representation of increase in complexity of (a) explicit, (b) implicit templates, and (c) their movement in the data table *(k* indicates delayed index)

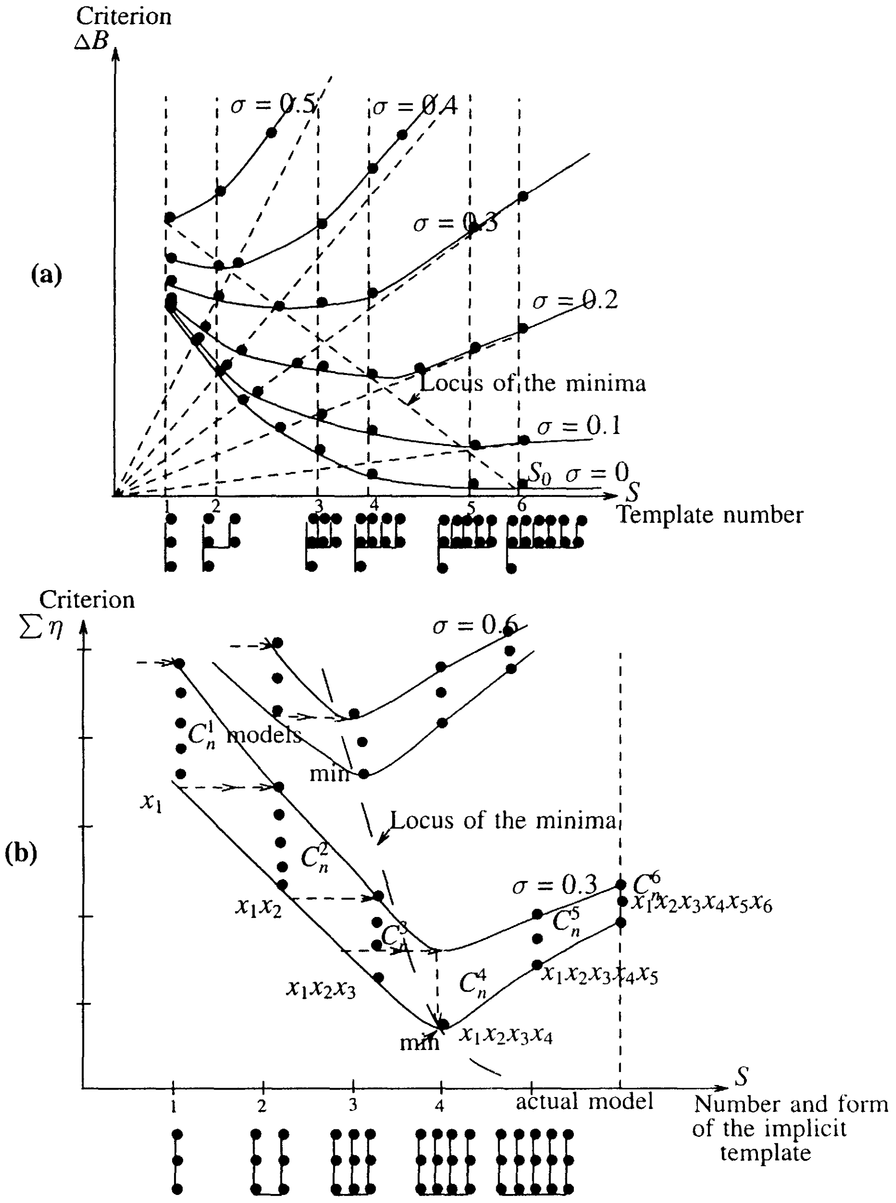


Figure 5.3. Results of experiments with (a) explicit patterns using vector models and (b) implicit patterns using objective systems analysis algorithm

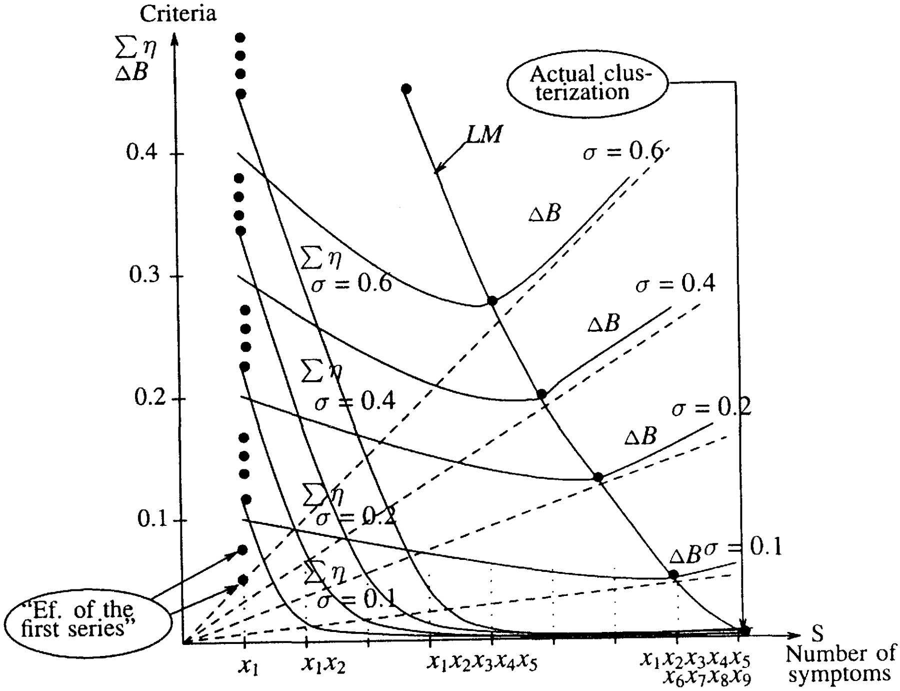


Figure 5.4. Results of experiments in clustering analysis, where LM stands for locus of the minima

Clusterization as investigation of a model in a "fuzzy" language

Clusterization algorithms differ according to their learning techniques that are categorized as learning "without teacher" and learning "with teacher." This means that in the latter case, the problem consists not only of the spontaneous division of the attribute space into clusters, but also of establishing the correspondence of each cluster with some point or region in the target function space. These algorithms are described for both the techniques as different stages "with teacher" and "without teacher." In other words, it leads to clusterization not only with the space of attributes *X* but also of the target function space *Y*, or of the united space *XY* where the target function is one of the attributes. As a result, clusterization $< X > \leftrightarrow < Y >$ or $< XY > \leftrightarrow < Y >$ is obtained—considered a certain "fuzzy" analogue of the model $y = f(\hat{x})$ of the object under investigation. The obtained model is optimal with respect to the criteria used and is unique for each object. In ideal data (without noise), it corresponds to the true target of the physical model. In noisy data, it corresponds to the nonphysical model—unique for that level of noise variance. Stability is considered according to the Darwin's classification of species and Mendelev's table of elements which confirm the uniqueness of classifications.

Artificial analogue of the target function

When the target function is not specified, it is sometimes necessary to visualize the output or target function through certain analysis. Visualization here means to make visible that which objectively exists but is concealed from a measurement process. This can refer to a person making a choice of initial data, not intentionally making it nonrepresentative, arranging it along certain axes—"weak-strong," "many-few," "good-bad," etc.—even when the target function is not completely known. A sample of conventionally obtained measurements thus contains information about the target function. Therefore all clusters must be represented in a sample for it to be representative. This is verified in various examples: in water quality problems, samples without any direct indication of the quality spanned the entire range from "purest" to "dirtiest" water. In tests of a person's intelligent quotient (IQ), it represents a broad range of values ($IQ = 10$ - 170). Since it is also determined by experts, it is always possible to check the idea of visualization of the target function. As results indicate, the experimentally measured target function correlates with its artificial analogue of correlation function (value ranges from 0.75 to 0.80), which is considered as adequate. Even for some experiments these are of higher values. The component analysis or Karhunen-Loeve transformation which is used to determine the analogue of the target function can be scalar, two-dimensional or three-dimensional (not more than three) corresponding to visualization of a scalar or a vector target function.

True, undercomplex, and overcomplex clusterizations

The view of clusterization as a model allows us to transfer the basic concepts and procedures of self-organization modeling theory into the self-organization theory of clusterization. A true clusterization corresponds to the so-called physical model which is unique and can be found in ideal and complete data using the first-level external criteria.

The consistent criterion expresses the requirement of clusterization structures as unbiased. Clusterization obtained using the set A must differ as little as possible from the clusterization obtained using the set B (A U B = W). The simplest among the unbiased (overcomplex) clusterizations is called true clusterization—the point with the optimum set of features denoted as "actual model" in Figure 5.3b. The overcomplex ones are located to the right of that point. Optimal clusterization corresponding to the minimum of the criterion is also unique, but only for a certain level of noise variance (the trivial consistent clusterization where the number of clusters is equal to the number of given points is not considered here). It is determined according to the objectives of the clusterization, and it cannot be specified. This explains the word "objective" in "objective computer clusterization." Optimal clusterizations are found by searching the set of candidate clusterizations differing from one another in the number of clusters and attribute ensembles. The first-level external criteria are explained previously in self-organization modeling. The basic criteria for clusterizations are defined analogously.

The consistency criterion of clusterizations is given as

$$\eta_c = (p - \Delta k)/p, \tag{5.1}$$

where p is the number of clusters or the number of individual points subject to clusterization in the subsets A and B; Δk is the number of identical clusters in A and B [70]. The regularity criterion of clusterizations is measured by the difference between the number of clusters (k_B) of the attribute space in the subset B and their actual number (k) indicated by the teacher. This is represented as $\Delta B = (k_B - k)$.

It has been established that in the problem of sorting models the values of the minimum-bias criterion depend on the design of the experiment and on the method of its partitioning into two equal parts. For an ideal data (without noise), the criterion is equal to zero both for the physical model and for all the overcomplicated models. The greater the difference between the separated sets A and B, the greater the value of the criterion. It is recommended that one can range the data points according to the variance of the output variable, then partition the series into equal parts of A and B. In clustering (delayed arguments are not considered), it is recommended that one choose a sufficiently small difference between the sets to preserve the characteristics of different clusters. If the clusters on the sets *A* and *B* are not similar, it is not worth using the consistent criterion. We cannot expect a complete coincidence of subsets *A* and *B*, which is inadmissible. Consequently, the problem of sorting clusters becomes a delicate one.

The consistent criterion is almost equal to zero for all the ensembles when the data are exact. It is recommended that the data be partitioned in such a manner that the criterion does not operate on the exact data. However, one can use various procedures to find the unique consistent cluster: (i) according to regularity criterion, (ii) according to system criterion of consistency $\sum \eta_c = \frac{1}{s}(\eta_{c_{(1)}} + \eta_{c_{(2)}} + \cdots + \eta_{c_{(s)}})$ by forming more supplementary consistent criteria computed on other *s* partitions, (iii) by adding noise to the data and from there finding the most noise-immune clustering, or (iv) by involving experts.

Necessity for regularization

Mathematical theory so far has not been able to suggest an expression for a consistency criterion indicating the closeness of all properties of models and clusterizations for the subsets *A* and *B*. The most widely used form of the criterion (minimum-bias criterion) stipulates the idea that the number of clusters ($k_A = k_B$) be equal and that there be no clusters containing different points ($\Delta k = 0$). The patterns of point divisions into *A* and *B* must coincide completely in the case of consistent clusterization. The consistent criterion is a criterion that is necessary but not sufficient to eliminate "false" clustrizations. This means that a circumstance might occur that leads to nonuniqueness of the selection. Several "false" clusterizations will be chosen along with the required consistent clusterizations. In these situations, regularization is necessary to filter out false clusterizations.

When the consistency criterion is used in sorting, a small number of clusterizations is found from which the most consistent one is selected—unique for each level of noise variance. For regularization, it is suggested that one use the consistent criterion once more, but employ a different method of forming it. To obtain a unique sample while sorting and using the consistency criterion, only a small number of clusterizations should be taken—chosen by an auxiliary unimodal criterion. Such an auxiliary, regularizing criterion is provided by a consistency criterion calculated on the other data sets *C* and *D*. For consistency of clusterizations, the patterns of point divisions into *A* and fi, as well as *C* and *D* must completely coincide. In addition to this, the optimal consistent clusterization must be unique. If more than one clusterization are obtained, then the regularization must be continued by introducing another two-subselections until a single answer is obtained. If the computer declares that there are no consistent clusterizations, then the sorting domain is extended by introducing new attributes and their covariances (higher order of the terms), introducing their values with delayed values in order to find a unique consistent clusterization.

High effectiveness of inductive algorithms

As in self-organization modeling, the model with optimal complexity does not coincide with the expert's opinions. The best cluster, being consistent and optimal according to the regularity precision, does not coincide with *a priori* specified expert decisions. Expert decisions are related to complete and exact data. The self-organization clustering that considers the effect of noise in the data, reduces the number of symptoms in the ensemble and the number of clusters. The greater the noise variance, the greater will be the reduction in the number. The computer takes the role of arbiter and judge in specific decisions concerning the results of modeling, predictions and clustering analysis of incomplete and noisy data. This explains the presence of the word "computer" in the name "objective computer clusterization."

It is simply amazing how much world-wide effort has been spent on building the most complex theories oriented toward, surely, the hopeless business of finding a physical model and its equivalent exact clusterizations by investigating only the domain of overcomplex structures. The revolution associated with the emergence of the inductive learning approach consists of the problem of identification of a physical model and clusterization. The problems of prediction are solved in the other direction—of proceeding from undercomplex biased estimates and structures. Optimal biased models and clusterizations are directly recommended for prediction. Advancements in this direction propose a procedure for plotting the "locus of the minima" (LM) of external criteria for identification of the physical model and true clusterization.

Calculation and extrapolation of locus of the minima

The analogy between the theory of self-organization modeling and the theory of self-organization clustering can be continued to find optimal undercomplex clusterizations. One can use either search for variants according to external criteria or calculation of the locus of the minima of these criteria.

The calculation and extrapolation of the locus of the minima of external criteria is an effective method of establishing true clusterization from noisy or incomplete data. A special procedure for extrapolating the locus of the minima or the use of the canonical form of the criterion is recommended in various works [138] and [45] for finding a physical model or an exact clusterization. (Refer to Chapter 3 for the procedures in case of ideal criteria.) One can only imagine the effect of the analytical calculation of the locus of the minima on various criteria. This is calculated for a number of values of the variance and for various distributions of perturbation probabilities.

Usage of canonical form of the criterion for extrapolating LM. All the quadratic criteria can be transformed into a normalized canonical form by dividing the trace of the matrix of the criterion. The criterion is expressed as follows.

$$CR = Y^T S_{0-m} Y, \tag{5.2}$$

where CR indicates an external criterion in the canonical form. Y and Y^T are the output vector and its transpose, correspondingly. S_{0-m} is the canonical matrix of the criterion for different structural complexities.

The mathematical expectation of the criterion for all the models is

$$\overline{CR} = Y^T S_{0-m} Y + \sigma^2 \, \mathrm{tr} S_{0-m}, \quad \text{where } S_{0-m} = \overline{S_0, S_m}. \tag{5.3}$$

For example, S_0 corresponds to a physical model, then

$$\frac{\overline{CR}}{\mathrm{tr} S_0} = \sigma^2, \tag{5.4}$$

and S_s corresponding to a nonphysical model, then

$$\frac{\overline{CR}}{\mathrm{tr} S_j} = \sigma^2 + \frac{Y^T S_j Y}{\mathrm{tr} S_j} = \sigma^2 + A, \quad A \geq 0. \tag{5.5}$$

Hence, $\frac{\overline{CR}}{\mathrm{tr} s_j} \geq \frac{\overline{CR}}{\mathrm{tr} s_0}$.

Theorem. The minimum of the mathematical expectation of the criterion in canonical form for nonphysical models is greater than it is for a physical model [138].

It is shown that all the criteria in canonical form create LM which coincides with the ordinate of the physical model (Figure 5.5). From a geometric point of view, transformation of the criterion to canonical form means rotation of the coordinate axes around the point S_0 and some small nonlinear transformation of the coordinate scale. Figure 5.5 exhibits the locus of the minima: (a) for an external criterion with the usual form and (b) for its canonical form taking the values of $\overline{CR}/$ tr$S$. This shows that with the use of the canonical form of the criteria, one can find a model in optimal complexity without adding any auxiliary noise to the data.

The choice of a rule for restoring the actual or physical model depends on the number of candidate models subject to descrimination, the perturbation level, and the type of criterion.

First rule. If the number of candidate models and the perturbation level are so small that the noise level σ^2 is not exceeded; there is no need for special procedures. The actual clustering is found by using the consistency criterion.

Second rule. If the number of models or candidate clusterings and the perturbation level are comparatively large, a "jump" to the left by the locus of the minima is observed (Figure 5.5a). By imposing supplementary noise on the data sample, one can find several points of the envelope of locus of the minima and use its extrapolation to determine the physical model or actual clustering [45].

Third rule. Addition of auxiliary noise is not needed if the criterion is transformed into canonical form. The ordinate of the minimum of the canonical criterion will indicate the optimal structure (or template) of the physical model or of the clustering if the perturbation variance is within considerable limits (Figure 5.5b).

Asymptotic theory of criteria and templates

In Chapter 3, we discussed the asymptotic properties of certain external criteria. For the mathematical expectation of the external criterion with an infinitely long data sample, the characteristic of the criterion-template sorting is unimodal which is required according to the principle of self-organization. One should not conclude from this result that every time-averaging of the criteria is well only in asymptotic behavior. But unimodality is attained considerably within the limits for a sample length of five to ten correlation intervals; however, a more accurate estimate of the required time-averaging of the criteria is to be found analytically—a subject of theoretical interest.

Asymptotic theory of templates is also not yet developed, although it has been established experimentally. The gradual increase in the number of models according to a specific template leads to an increase in the probability (number of occurrences) of attaining unimodality. Figure 5.6 demonstrates the proposed dependence using the consistency criterion in the plane of "perturbation variants-template complexity."

The future asymptotic theory of templates requires the investigation of the behavior not of the average line of criterion variation, as one selects out of each cluster of feature variants that comes for sorting only one model—the best. This is done by distinguishing among the patterns of variation using a partial, solitary, and overall consistency criteria. For features with noiseless data in clusterizations, the partial nonoverall consistency criterion is identically equal to zero for the entire duration of sorting if the subsamples A and B are close to each other, but nonetheless distinct. The interval of the zero values of the consistency criterion shrinks with sufficiently high probability as the perturbation variance

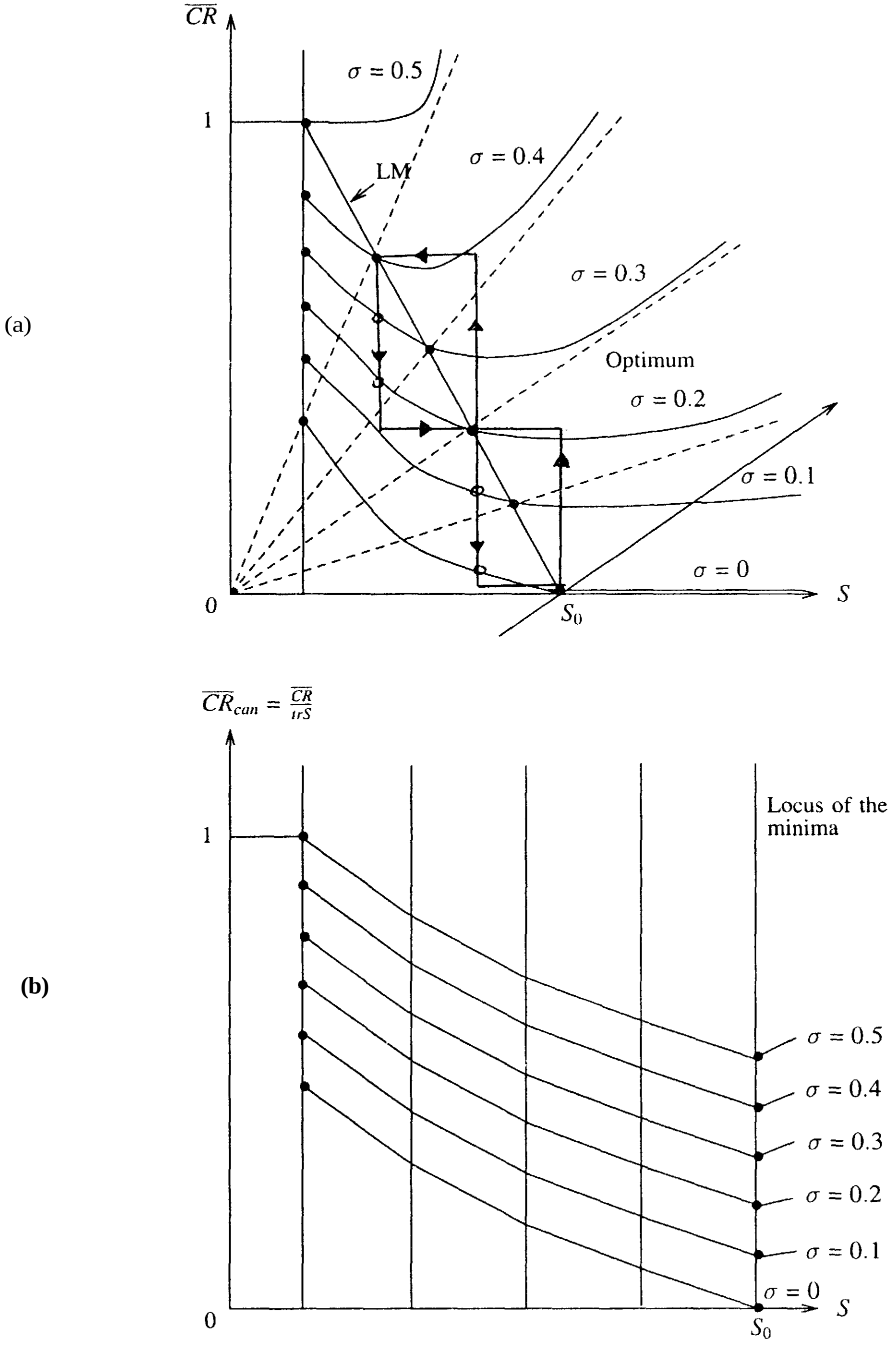


Figure 5.5. Locus of minima (LM) in transition (a) to the ordinary, and (b) to the canonical forms of the criteria, depending on model complexity *S* and noise dispersion *a*.

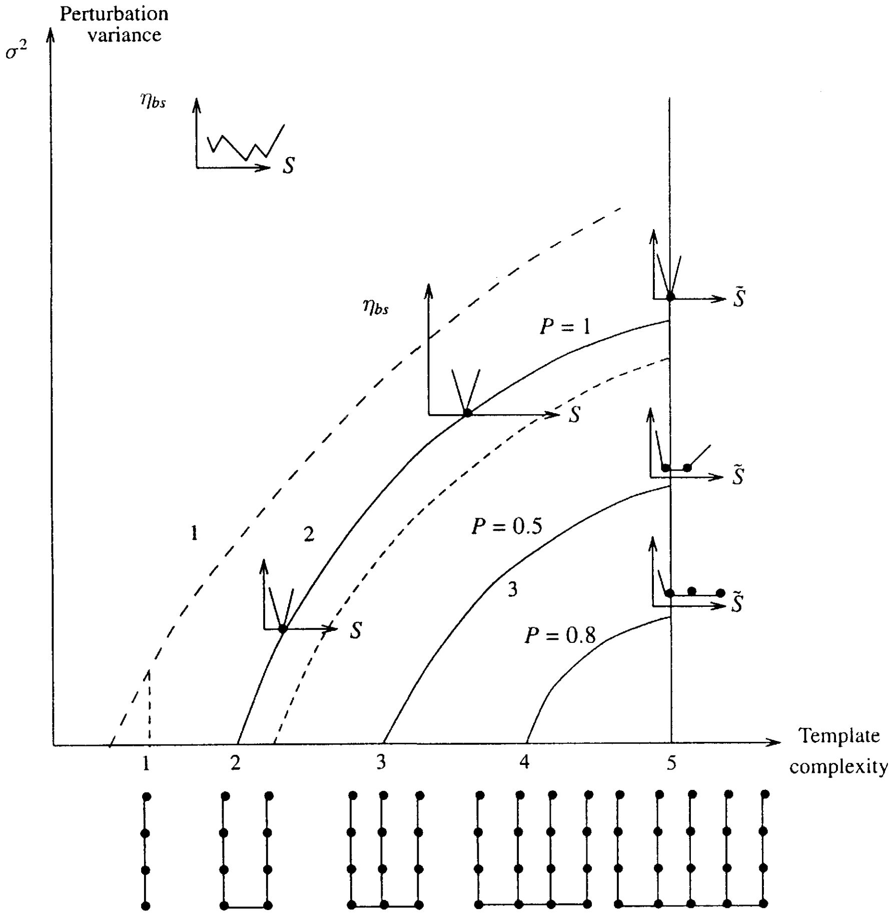


Figure 5.6. Proposed change in probability P of attaining unimodality of the consistency criterion: (1) region of loss of unimodality, (2) region of unimodality without extension of determination, (3) region in which extension of determination required

increases. When it becomes sufficiently small to distinguish between the templates, it becomes expedient to extend the sorting by using an accuracy criterion or a series of consistency criteria calculated for various partitions of data sample. For a larger perturbation variance, it will be in the region of unimodality of a solitary criterion, where a larger perturbation variance is required for more complex templates. Strictly speaking, this serves as the basis for the asymptotic theory of templates. For excessively large perturbations, it becomes impossible to find an optimal consistent model or clusterization, since the regular nature of the curve disappears (Figure 5.6).

2 METHODS OF SELF-ORGANIZATION CLUSTERING

Unlike the sorting of partial models, which is almost always obtained, the sorting of clusters can be implemented only for a sufficiently large number of points that are located favorably in the symptoms (variables). The importance of special experimental designs are enhanced in this section.

If there are m symptoms, one can construct 2^m different ensembles and evaluate them by a suitable external criterion; for example, regularity criterion for an accurate approach and the system criterion of consistency for a robust approach. This corresponds to unsupervised learning because of the absence of specific objectives. If the objective is specified as the ensembles are grouped to a known target function, then it corresponds to supervised learning. The self-organization clustering methods vary according to the techniques used for the reduction of computational volume.

The *first method* is a selection-type of sorting method based on unsupervised learning [39]. At the first step, all the symptoms at the time of succession are evaluated by the specified basic criterion and the best of F (freedom-of-choice) are chosen (for example, $F = 3$ and the symptoms are x_1, x_7 and x_9). At the second step, all the ensembles that contain two symptoms are evaluated. These ensembles include all the symptoms selected at the first step.

$$\begin{array}{c|c|c|c|c|c|} - & x_1x_2 & x_1x_3 & x_1x_4 & x_1x_5 & x_1x_6 \\ x_7x_1 & x_7x_2 & x_7x_3 & x_7x_4 & x_7x_5 & x_7x_6 \\ x_9x_1 & x_9x_2 & x_9x_3 & x_9x_4 & x_9x_5 & x_9x_6 \end{array} \cdots \qquad (5.6)$$

The F best ensembles (for example, $F = 3$, and they are x_1x_7, x_3x_7, and x_1x_4) are selected. At the third step, the ensembles that have three symptoms by including the ensembles selected at the second step are evaluated. This evaluation continues until the 3 x m ensembles are selected.

The *second method*, which is based on correlation analysis [70], is suitable for the precision in the approach. Here, one can obtain a series of m symptoms which range according to their effectiveness; only m different ensembles are evaluated by the criterion.

The *third method* uses one of the basic inductive learning algorithms, either combinatorial or multi-layer, to find m effective ensembles. For example, one can use a device like combinatorial type of "structure of functions" for generating all combinations of ensembles by limiting the number of symptoms. The consistent criterion is used with the data sequences of A and B that are close to each other.

The latter two methods correspond to the supervised learning (learning with teacher) because they use information about the output vector Y based on the comparison among the actual and the estimated data. One way of doing this is by specifying the output data from the experiment and another way is by using the orthogonal Karhunen-Loeve projection method for obtaining the artificial data.

The above methods does not limit the scope of all possibilities. They are feasible only when the unimodality characteristic of the "criterion-clustering complexity" is ensured. These we see in detail below.

2.1 Objective clustering—case of unsupervised learning

There are various computer algorithms that have been proposed for separating a set of ensembles or clusters given in a multidimensional space of variables or symptoms. This includes the classical algorithm of ISODATA (Iterative Self-Organizing Data Analysis Techniques Algorithm) [124] that is based on comparing all possible clusters using the minimum distance criterion. In this program, the number of clusters are specified in advance by the expert.

Objective clustering is envisaged by the inductive approach in which a gradual increase in the number of clusters is specified to the computer and are compared according to the consistent criterion. In separating a multidimensional data space into clusters, the consistent criterion may, for example, stipulate that the partitioned clusters differ from one another as little as possible as they are partitioned according to the odd and even-indexed points of initial data. As is well known, typographical images of some pattern consist of dots. Even when the even or odd dots are excluded, it preserves the image with large numbers of initial data points. If the original image is chaotic; i.e., even if it contains no information conforming to some law, the criterion allows discovery of a physical law.

The object or image is given in a multidimensional space represented in the form of observation data with symptoms $x_1, x_2, \cdots, x_m$. The first part of the problem consists of dividing the space into a specified number of regions or clusters using the measurements of distance between the points [124]. The number of clusters is specified in advance by the experts. Self-organization involves iteration of such clusterings for various numbers of clusters from $k = 2$ to $k = N/2$, where N is the number of data points. It also invloves comparison of results by the consistent criterion—non-contradictory clusters are selected. A single-valued choice is achieved by regularization. Here regularization is selecting the single most appropriate cluster from several non-contradictory clusterings indicated by the computer. The role of regularization criterion is to use the minimizing function which takes into account the number of k and number of variables or symptoms m according to the computer's and expert's clusterings.

$$\rho = [(k_{exp} - k_{comp})^2 + (m_{exp} - m_{comp})^2], \tag{5.7}$$

where k_{exp} is the number of clusters specified by the expert and k_{comp} is the number of clusters in the process of computer clustering.

If k_{exp} is known, then the computer completes the determination of clusters—for example, by using the function $L = k/m$. This is also determined by other relations, in case it is required by agreeing results on three equal parts of the selection.

Even if the k_{exp} is not known, one can use the consistent criterion calculated in other parts of the data sample. It evaluates the degree of non-contradiction on various clusters and helps to choose the best one.

Example 1. Clustering of water quality indices (one-dimensional problem).

The initial data contain the following variables: x_1—suspended matter in mg/liter, x_2—chemical consumption of oxygen (CCO), x_3—mineralization in mg/liter, x_4—carbohydrates in mg/liter, and x_5—sulphates in mg/liter. The data is normalized according to the formula $x_{i_{norm}} = \frac{x_i - x_{i_{min}}}{x_{i_{max}} - x_{i_{min}}}$. The measurements are averaged on seven years of data for each station. The data sets A and B include all stations with even and odd numbers, respectively.

The algorithm is confronted with the problem of isolating all non-contradictory clusterings using the given set of variables and all subsets which could be obtained from them. Thus, the water quality expert could choose the most valid clustering and find the number of clusters and the set of variables that are optimal under given conditions. It computes the value of the criterion for all possible combinations of the set of given variables. In this case the validity of clustering is not verified because of the absence of expert clustering. The sorting process showed that it is not possible to obtain a non-contradictory cluster using all five variables. For each identified cluster, the centers and boundaries are found and the water quality at the given station using the corresponding variables from the cluster is computed.

Example 2. Clustering of water quality along the series of water stations along a river system.

In this case, expert clustering is known. It is established based on the information available on ecologic-sanitary classification of the quality of surface waters of dry land. It differs from certain variables which are absent from the data (out of total of 21 variables, only 14 participated in the example). The data of 14 variables is normalized and separated into two sets A and B.

The number of clusters specified by experts is $k = 9$ with the variables $m = 14$. There is no single set of variables chosen from the given 14 variables which would yield a non-contradictory partition of the stations into nine clusters as required by the experts. This means that the expert cluster is contradictory.

Non-contradictory partitions into eight clusters are given by a comparatively small number of variables which include x_{14}, x_6x_{14}, $x_2x_7x_{10}x_{14}$ and $x_2x_4x_6x_{10}x_{14}$. Many sets of variables give non-contradictory partitions into seven clusters; eight such sets are x_1x_2, x_1x_4, x_1x_5, x_1x_6, x_1x_7, x_1x_{12}, x_1x_{14}, x_4x_{13}, and 22 sets—each having three variables (from $x_1x_2x_4$ to $x_1x_{12}x_{14}$). The following three sets each with 10 variables give a partition which is closest to one of the expert's clusterings:

$$x_1x_2x_3x_4x_5x_6x_7x_9x_{10}x_{12},$$

$$x_1x_2x_3x_4x_5x_6x_7x_9x_{10}x_{14},$$

$$x_1x_2x_3x_4x_5x_6x_7x_9x_{12}x_{14}.$$

The sets with higher number of variables (11, 12, 13 and 14) do not increase the number of clusters. The set of variables $m = 9$ is denoted as optimal in this example which gives a non-contradictory partition into seven clusters. The boundaries, the stations making up their composition, and the cluster centers are indicated for all non-contradictory clusters for further analysis of water quality.

2.2 Objective clustering—case of supervised learning

Classification, recognition, and clusterization of classes are similar names given for processing a measured input data. The space of measured data for input attributes $X(x_1, x_2, \cdots, x_m)$ with a given space of output $Y(y_1, y_2, \cdots, y_l)$ representing a target or goal function (where $l \leq m$) is common in these algorithms. The problem task is to divide both spaces into certain subspaces or clusters to establish a correspondence between the clusters of the attribute space and goal function space $X \leftrightarrow Y$.

Unlike in traditional subjective algorithms, the number of clusters are not specified in advance in objective clustering, but the number of clusters is chosen by the computer so that clusterization is consistent. This means that it remains the same in different parts of the initial input data. This number is reduced to preserve the consistency in case of noisy and incomplete data.

As it is mentioned earlier, the objective computer clustering is based on the search for the variants of ensemble of attributes and the number of clusters using the consistency criterion on the given measured data assuming certain errors. The algorithm gives the consistent clusterizations while all existing measurements are distributed over the clusters. The new measurements that do not participate in the clustering also belong to certain cluster, according to the nearest neighbor rule, or according to the minimum-distance rule from the center of the cluster.

The search for the attribute ensembles and for the number of clusters leads to multiple solutions: several variants of ensembles giving consistent clusterizations are found on the plane "ensemble of attributes-number of clusters." This is solved by further determination of consistent clusterings using some second-level criterion or by inquiring from experts.

Table 5.1. Initial Data

No.	x_1	x_2	x_3	x_4	x_5	y
1	2.131	10.41	69.22	73.52	4.43	12.23
2	2.031	9.797	69.26	74.10	4.84	11.86
3	2.076	9.892	69.06	73.42	4.36	11.72
4	2.084	10.09	69.02	73.36	4.34	11.83
5	2.057	9.816	68.97	73.32	4.45	11.47
.	.	.	.	.	.	.
.	.	.	.	.	.	.
19	2.109	10.05	68.81	73.16	4.31	12.05
20	2.143	10.52	68.76	73.01	4.25	12.48
21	2.115	10.24	68.77	73.07	4.30	12.22
22	2.150	10.45	68.71	73.10	4.39	12.38
23	1.919	9.295	68.66	73.06	4.40	10.96
24	2.046	9.840	68.63	73.06	4.43	11.64
.	.	.	.	.	.	.
.	.	.	.	.	.	.
37	2.005	9.631	68.01	72.33	4.32	11.50
38	2.047	9.937	68.06	72.43	4.37	11.67
39	2.013	9.864	68.06	72.42	4.36	11.60
40	2.123	10.37	68.03	72.42	4.39	12.30

Example 3. Objective clustering of the process of rolling of tubes [71].

Here the problem of objective partitioning of an m-dimensional space of features x_1, x_2, • • ·, x_m into clusters corresponding to compact groups of images is considered; each image is defined by a data sample of observations.

Objective clustering of images (data points) is done based on sorting a set of candidate clusterings using the consistency criterion to choose the optimally consistent clusterings. The data is divided into four subsets: $A \cap B$ and $C \cap D$. Here the concept of dipoles (pairs of points close to each other) is used; one vertex of a dipole goes into one subsample and the other into another. Thus, the greatest possible closeness of points forming the subsamples is achieved. This example demonstrates the various stages of self-organization clustering algorithm which does not require computations of the mean square distances between the points.

The table of initial data is given (Table 5.1), where x_1 is the length of the blank, x_2 is the length of the tube after the first pass, x_3 and x_4 are the distances between the rollers in front of the two passes, x_5 (= x_4 — x_3) is the change in distance between the rollers, and $y = f(x_1, x_2, \cdots, x_5)$ is the length of the tube.

The objective clustering is conducted in the five-dimensional space of the features $x_1, x_2, \cdots, x_5$. The clustering for which we obtain the deepest minimum of the consistency criterion is the optimal one. The stage-wise analysis of the algorithm is shown below.

Stage 1. To compute the table of interpoint distances. The first $N = 34$ data points from the 40 points of the original sample are used to form the subsets $A \cap B$ and $C \cap D$. The remaining six points are kept as testing sample to check the final results of clustering and for establishing the connection between the output variable y and the cluster numbers. The initial data table is represented as a matrix $X = [x_{ij}]$; $i = 1, 2, \cdots, N$ and $j = 1, 2, \cdots, m$ (here $N = 34$ and $m = 5$).

Table 5.2. Interpoint distances between dipoles

No.	1	2	3	4	5	6	...	32	33	34
1	0	1.015	0.310	0.171	0.484	0.344	...	3.779	1.952	3.484
2		0	0.743	0.943	0.845	1.434	...	5.211	2.339	5.376
3			0	0.449	0.0325	0.399	...	2.547	1.195	2.561
4				0	0.092	0.636	...	2.503	1.150	2.391
5					0	0.318	...	2.115	0.895	2.169
6						0	...	1.990	1.465	2.361
.							.	.	.	.
.							.	.	.	.
.							.	.	.	.
32								0	0.966	0.111
33									0	0.954
34										0

The interpoint distances are calculated as

$$d_{ik} = \sum_{j=1}^{m} (x_{ij} - x_{kj})^2, \quad i = \overline{1,N}; \ k = \overline{i+1,N}. \tag{5.8}$$

The results are shown in the Table 5.2.

Stage 2. To determine the pairs of closest points and partition into subsets. The clusterings are to be identified in the two subsets of $A \cap B$ and $C \cap D$. Thus, the coincidence of clusters is required, indicating that they are consistent. This leads to the attainment of a unique choice of consistent clustering.

The subsets A n B and C n D are formed using the values of the dipoles. The dipoles are arranged in increasing length: for N = 34, there are $N(N$ — 1)/2 = 561 dipoles. The shortest dipoles are exhibited as

1) 11 0.0020 14, 2) 12 0.0038 13, 3) 23 0.0850 25, ...

To form the subsets A and B, the first $(\frac{N}{2} - 1) = 16$ shortest dipoles are chosen in such a way that the data points are not repeated. In this specific example, it turns out that these 16 dipoles are obtained from the first 389 dipoles; the 17th dipole which satisfies the condition is obtained at the end of the series; i.e., the 561st dipole connects the points 2 and 34 at a length of $d_{2,34}$ = 5.376 units.

The following 16 shortest dipoles belong to the subsamples A and B.

1) 11 – 14 2) 12 – 13 3) 23 – 25 4) 26 – 27
5) 16 – 19 6) 10 – 15 7) 5 – 8 8) 17 – 24
9) 20 – 22 10) 3 – 7 11) 31 – 34 12) 29 – 33
13) 6 – 18 14) 9 – 21 15) 1 – 4 16) 28 – 30

From the remaining dipoles, the 16 shortest dipoles are chosen in an analogous manner to form the subsamples C and D.

1) 18 – 23 2) 13 – 21 3) 16 – 17 4) 8 – 10
5) 14 – 15 6) 12 – 19 7) 3 – 5 8) 9 – 22
9) 30 – 31 10) 11 – 24 11) 4 – 7 12) 32 – 34
13) 20 – 27 14) 6 – 25 15) 26 – 33 16) 1 – 2

The dipoles obtained in this way enable the formation of the set of points into the subsets A, fl, C, and *D.*

$$\begin{array}{ll}
A: & 11,12,25,26,16,15,8,24,20,7,34,29,18,21,4,30;\\
B: & 14,13,23,27,19,10,5,17,22,3,31,33,6,9,1,28;\\
C: & 23,21,16,10,14,19,5,22,31,24,7,34,27,25,33,1;\\
D: & 18,13,17,8,15,12,3,9,30,11,4,32,20,6,26,2.
\end{array}$$

Stage 3 To sort the clusterings according to the consistency criterion.
The following steps are followed:

1. *Grouping the subsets into 16 clusters (k -* 16). The points in subsets *A* and *B* are indexed from 1 to 16 as vertex numbers, indicating a group of 16 clusters shown below:

$$\begin{array}{c} A \\ k=16 \\ B \end{array}\left\{\begin{array}{cccccccccccccccc}
11 & 12 & 25 & 26 & 16 & 15 & 8 & 24 & 20 & 7 & 34 & 29 & 18 & 21 & 4 & 30\\
\dot{1} & \dot{2} & \dot{3} & \dot{4} & \dot{5} & \dot{6} & \dot{7} & \dot{8} & \dot{9} & \dot{10} & \dot{11} & \dot{12} & \dot{13} & \dot{14} & \dot{15} & \dot{16}\\
14 & 13 & 23 & 27 & 19 & 10 & 5 & 17 & 22 & 3 & 31 & 33 & 6 & 9 & 1 & 28\\
\dot{1} & \dot{2} & \dot{3} & \dot{4} & \dot{5} & \dot{6} & \dot{7} & \dot{8} & \dot{9} & \dot{10} & \dot{11} & \dot{12} & \dot{13} & \dot{14} & \dot{15} & \dot{16}.
\end{array}\right.$$

Number of corresponding vertices or clusters:

$$\Delta k = 1+1+1+1+1+1+1+1+1+1+1+1+1+1+1+1 = 16.$$

In each subset A or *B,* the upper row denotes the actual data point and the lower row denotes the number of the vertex of the dipole. If the number of the vertices coincide, then those vertices are called "corresponding" vertices. Here, all vertices of subset A correspond to the vertices of the subset *B.* The consistency criterion is computed as $\eta_c = (p— \Delta k)/p = (16 — 16)/16 = 0$, where *p* is considered the total number of vertices and Δk is the corresponding vertices which coincide.

2. *Grouping the subsets into 15 clusters (k -* 15). Tables of interpoint distances are to be compiled for the points of each subset A and *B* (Tables 5.3 and 5.4, correspondingly). Points 2-14 in subset A and points 1-8 in subset *B* are the closest to each other.

   For the evaluation of the consistency criterion, it is grouped into 15 clusters in the following form.

$$\begin{array}{c} A \\ k=15 \\ B \end{array}\left\{\begin{array}{l}
\dot{1}\ \boxed{\begin{matrix}\dot{2}\\14\end{matrix}}\ \dot{3}\ \dot{4}\ \dot{5}\ \dot{6}\ \dot{7}\ \dot{8}\ \dot{9}\ \dot{10}\ \dot{11}\ \dot{12}\ \dot{13}\ \boxed{\begin{matrix}\dot{2}\\14\end{matrix}}\ \dot{15}\ \dot{16}\\
\boxed{\begin{matrix}\dot{1}\\8\end{matrix}}\ \dot{2}\ \dot{3}\ \dot{4}\ \dot{5}\ \dot{6}\ \dot{7}\ \boxed{\begin{matrix}\dot{1}\\8\end{matrix}}\ \dot{9}\ \dot{10}\ \dot{11}\ \dot{12}\ \dot{13}\ \dot{14}\ \dot{15}\ \dot{16}.
\end{array}\right.$$

Number of corresponding vertices:

$$\Delta k = 0+0+1+1+1+1+1+0+1+1+1+1+1+0+1+1 = 12.$$

The double number of the vertices indicate the formation of a cluster consisting of two points. Having the corresponding vertices as $\Delta k = 12$, the consistency criterion is $\eta_c = (16 - 12)/16 = 0.25$.

3. *Grouping the subsets into 14 clusters (k =* 14). Again the tables of interpoint distances are compiled, considering the formed clusters from the previous step. According to the nearest neighbor method, the distance from a cluster to a point is taken to be the

Table 5.3. Interpoint distances for subset A

No.	11	12	25	26	16	15	8	24	20	7	34	29	18	21	4	30
11	0	0.184	0.363	0.219	0.098	0.032	0.044	0.067	0.572	0.107	1.510	0.929	0.290	0.221	0.193	0.952
12		0	1.021	0.141	0.047	0.189	0.216	0.216	0.142	0.238	1.766	1.213	0.915	0.022	0.077	1.028
25			0	0.782	0.694	0.343	0.429	0.358	1.697	0.614	1.424	0.711	0.026	1.047	1.005	1.222
26				0	0.052	0.205	0.342	0.093	0.223	0.465	0.956	0.600	0.791	0.075	0.359	0.425
16					0	0.081	0.151	0.065	0.248	0.225	1.385	0.803	0.637	0.046	0.145	0.740
15						0	0.039	0.047	0.597	0.135	1.586	0.782	0.301	0.234	0.179	0.965
8							0	0.139	0.683	0.042	1.934	1.119	0.307	0.295	0.130	1.283
24								0	0.524	0.264	1.109	0.521	0.348	0.201	0.314	0.598
20									0	0.716	1.649	1.467	1.631	0.085	0.386	0.939
7										0	2.298	1.417	0.451	0.346	0.108	1.568
34											0	0.671	1.726	1.461	2.391	0.136
29												0	0.897	1.078	1.519	1.376
18													0	0.979	0.827	1.337
21														0	0.171	0.799
4															0	1.544
30																0

Table 5.4. Interpoint distances for subset B

No.	14	13	23	27	19	10	5	17	22	3	31	33	6	9	1	28
14	0	0.173	0.315	0.173	0.065	0.048	0.078	0.027	0.446	0.152	1.156	0.739	0.296	0.623	0.685	1.021
13		0	0.938	0.108	0.029	0.305	0.214	0.102	0.092	0.208	1.404	0.918	0.849	0.144	0.292	0.946
23			0	0.734	0.647	0.243	0.455	0.438	1.392	0.672	1.156	0.802	0.185	1.764	1.823	1.546
27				0	0.081	0.338	0.334	0.099	0.134	0.452	0.738	0.533	0.900	0.291	0.746	0.522
19					0	0.167	0.128	0.028	0.182	0.158	1.220	0.765	0.608	0.288	0.427	0.924
10						0	0.034	0.077	0.640	0.125	1.412	0.723	0.218	0.841	0.761	1.362
5							0	0.084	0.531	0.0325	1.718	0.895	0.318	0.663	0.484	1.503
17								0	0.288	0.160	1.099	0.580	0.475	0.451	0.590	0.958
22									0	0.542	1.296	0.862	1.442	0.036	0.446	0.801
3										0	2.086	1.195	0.399	0.612	0.310	1.745
31											0	0.606	1.966	1.729	2.968	0.410
33												0	1.465	1.231	1.952	1.108
6													0	1.691	0.344	2.014
9														0	0.323	1.040
1															0	2.117
28																0

smaller of the two distances. For example, the distance from point 1 to cluster 2,14 is the smaller of the two quantities d_{1-2} = 0.184 and d_{1-14} = 0.221; ie., $d_{1-2,14}$ = 0.184. Thus, the closest points to each other are 3-13 (subset *A)* and 5-1,8 (subset *B*).

The third candidate is grouped into 14 clusters of the form

$$\begin{array}{cl} \begin{matrix} A \\ k = 14 \\ B \end{matrix} & \left\{ \begin{array}{l} \dot{1}\ \ \boxed{\begin{matrix}\dot{2}\\14\end{matrix}}\ \boxed{\begin{matrix}\dot{3}\\13\end{matrix}}\ \dot{4}\ \ \dot{5}\ \ \dot{6}\ \dot{7}\ \ \dot{8}\ \ \dot{9}\ \ \dot{10}\ \ \dot{11}\ \ \dot{12}\ \boxed{\begin{matrix}\dot{3}\\13\end{matrix}}\ \boxed{\begin{matrix}\dot{2}\\14\end{matrix}}\ \ \dot{15}\ \ \dot{16} \\ \boxed{\begin{matrix}\dot{1}\\5\\8\end{matrix}}\ \ \dot{2}\ \ \ \dot{3}\ \ \dot{4}\ \boxed{\begin{matrix}\dot{1}\\5\\8\end{matrix}}\ \dot{6}\ \dot{7}\ \boxed{\begin{matrix}\dot{1}\\5\\8\end{matrix}}\ \dot{9}\ \ \dot{10}\ \ \dot{11}\ \ \dot{12}\ \ \dot{13}\ \ \ \dot{14}\ \ \dot{15}\ \ \dot{16}. \end{array} \right. \end{array}$$

Number of corresponding vertices:

$$\Delta k = 0 + 0 + 0 + 1 + 0 + 1 + 1 + 0 + 1 + 1 + 1 + 1 + 0 + 0 + 1 + 1 = 9$$

and $\eta_c = (16 - 9)/16 = 0.437$.

4. *Fourth and subsequent steps.* Continuation of the partitioning of the subsets into clusters and evaluation by consistency criterion is followed from k = 13 to k = 2.

   For the last two clusterings; i.e., in case of k = 2, η_c = (16 — 16)/16 = 0, and in case of k = 3, η_c = (16 - 16)/16 = 0.

All groupings of the clusterings is complete. From the above evaluation, the consistent clusterings for k = 2, 3, and 16 can be chosen because η_c = 0 in these groupings.

One can note that if the table of interpoint distances consists of two equal numbers, then the number of clusters changes by two units. To avoid this, one must either raise the accuracy of the measurement distances in such a way that there will not be equal numbers in the table, or skip the given step of sorting of clusterings in one of the subsets. The consistency criterion is used only when the number of clusters is the same on two subsets $A \cap B$ and $C \cap D$; otherwise, the amount of sorting increases and it ends up with bad results.

To reduce the computational time of the algorithm, the comparison of the variants of the clusterings can be started with eight clusters instead of 16 clusters. This means that at the first step the points are not combined by two, but by eight points.

Stage 4. Repetition of clustering analysis on subsets *A* and *B* for all possible sets of variable attributes (scales) and compilation of the resulting charts (Figure 5.7a).

The cluster analysis described above should be repeated for all possible compositions of the variable attributes. As there are m = 5 attributes, there are altogether 2^5 — 1 = 31 variants. The dots in the figure indicate the most consistent clusterings which are obtained on the subsets *A* and *B.*

Stage 5. To single out the unique consistent clustering with the aid of experts or by using the subsets *C* and *D* (regularization).

It is desirable to choose a single most consistent one from the clusterings obtained on the subsets *A* and *B.* This can be done in two ways: One way of singling out is with the help of experts for whom examination of a small number of variants of clusterings does not constitute any great difficulty. The unique clustering suggested by the expert might not be the most consistent clustering, but merely one of the sufficiently consistent clustering. Another way is by repeating the clustering analysis on subsets *C* and *D* to obtain a clustering

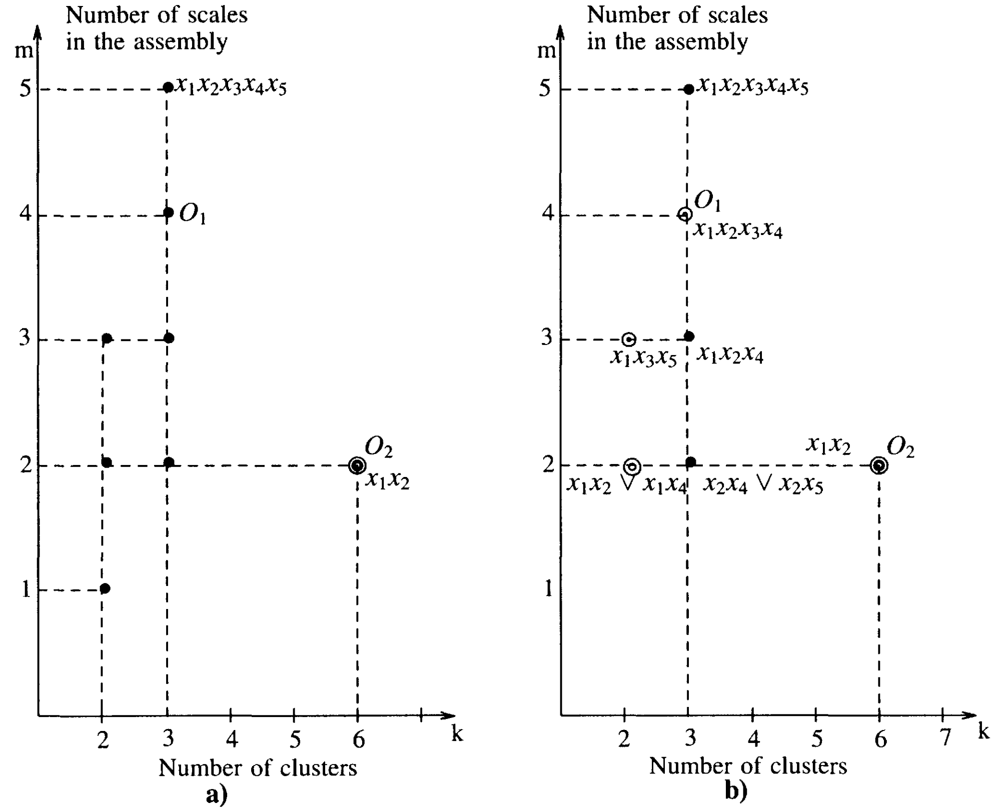


Figure 5.7. Results of search for the most consistent clusterings on (a) subsamples A and B and (b) subsamples C and D

that will prove to be sufficiently consistent both for the subsets A n B and C n D. Figure 5.7b shows the results of choice of consistent clusterings on subsets C and D. The value of the consistency criterion for the clustering corresponding to the point O_2 is zero both on the subsets A n B and $C \cap D$. For the clustering O_1, it is zero only for C n D. Here clustering 02 is considered to be the true most consistent ones.

If unique clustering is not obtained, the points are further divided into three equal subsets, thus forming another consistency criterion and so on until the goal of the regularization—a single consistent clustering—is achieved.

Figure 5.7 shows less than eight clusters (out of the 16 possible ones) along the abscissa, since further increase in their number yields an inadmissibly small mean number of points in each of them (total 34 points are subjected to grouping in clusters).

For *reducing the sorting of attributes*, it is recommended that

1. the attribute sets for which half or more of the dipoles on A n B (or C n D) do not coincide are not considered, and
2. for analysis on subsets C and D, one considers only those attribute sets for which small values of the criterion during the analysis on the subsets A and B are obtained.

Stage 6. Results of the two clusterings corresponding to O_1 and O_2.

Corresponding to the point O_1, three clusters are obtained with respect to four scales of attributes x_1, x_2, x_3, and x_4. The points of the original data sample are distributed among the clusters as below (the point numbers and the mean values of the output variable y are given):

$$\begin{aligned}
&\text{1st cluster}: && 6, 18, 23, 25;\ \bar{y} = 10.99m;\\
&\text{2nd cluster}: && 30, 31, 32, 34;\ \bar{y} = 11.58m;\\
&\text{3rd cluster}: && 1, 2, 3, 4, 5, 7, 8, 9, 10, 11, 12, 13, 14, 15,\\
& && 16, 17, 19, 20, 21, 22, 24, 26, 27, 28, 29, 33;\ \bar{y} = 11.799m.
\end{aligned}$$

Corresponding to the point O_2, six clusters are obtained with respect to the two scales of attributes x\ and x_2.

$$\begin{aligned}
&\text{1st cluster}: && 6, 18, 23, 25;\ \bar{y} = 10.99m;\\
&\text{2nd cluster}: && 29, 32;\ \bar{y} = 11.20m;\\
&\text{3rd cluster}: && 1, 2, 3, 5, 7, 8, 9, 10, 11, 14, 15, 17, 24, 31, 33, 34;\ \bar{y} = 11.47m;\\
&\text{4th cluster}: && 4, 16, 19, 26, 27, 30;\ \bar{y} = 11.83m;\\
&\text{5th cluster}: && 12, 13, 21, 28;\ \bar{y} = 11.93m;\\
&\text{6th cluster}: && 20, 22;\ \bar{y} = 12.43m.
\end{aligned}$$

Stage 7. To check the optimal clustering using the checking sample of data points (35 to 40) according to the prediction accuracy of required quality of the tube length.

The single consistent clustering can be used to predict the output variable y from the cluster number. For example, let us consider the three clusters corresponding to the point O\ with the attributes x_1, x_2, x_3, and x_4 (the three clusters with the point numbers and mean values of the variable y are given above). The mean values of y are arranged in an increasing order and the regression line for y according to the groupings of clusters N is given in Figure 5.8. A new point belongs to the cluster for which the distance from it to the closest point of the cluster is least; knowing the cluster, the estimated value of y can be obtained from the figure. This type of prediction is checked for the testing sample points 35 to 40. Out of six points, five are correctly predicted.

2.3 Unimodality—"criterion-clustering complexity"

We understand that the experimental design is feasible only when the unimodality of the "criterion-clustering complexity" characteristic is ensured. This can be done in three ways to determine the optimal consistent clustering: (i) extend the cluster analysis using a regularity criterion for further precision, (ii) design the cluster analysis for using a overall or system criterion of consistency by increasing the number of summed partial consistency criteria, and (iii) design the experiment by applying a supplementary noise to the data.

The applicability of the first method is demonstrated in the preceding example.

The second method of attaining unimodality is when an increase in the number of partial criteria which constitute the overall consistency criterion reduces the number of consistent clusterings from which an optimal one is to be selected. Specially designing the experiment can make this method very efficient in yielding a single consistent clusterization. The following example demonstrates the usefulness of this method.

Example 4. Investigation of the consistent criterion by computational experiments [69].

Here is a test example to clarify whether (i) it is possible to select a data sample such

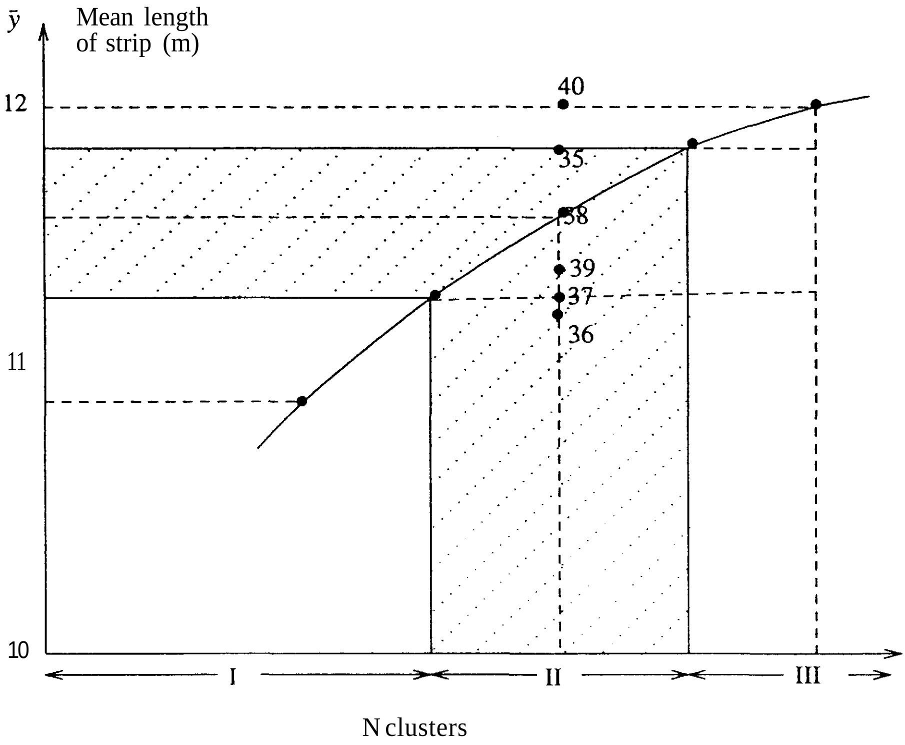


Figure 5.8. Regression line for prediction of mean strip length for the cluster number N for the set of x_1, x_2, x_3, x_4

that sorting of clusterings by the consistency criterion yields a unique solution and (ii) the overall consistency criterion leads to a unique solution.

The consistency criterion is expressed as $\eta_c = (k — \Delta k)/k$, where k is the number of clusters and Δk is the number of identical clusters in the subsets A and B.

According to the procedure involved in the experimental design of cluster analysis, the original data sample is divided into two equal parts by ranking their distances from the coordinate origin. Then the consistent clusterings are found by complete sorting of hypotheses about the number of clusters, proceeding from $k = N/2$ to a single cluster, where N is the total number of points in the data sample. The initial data sample along with their ranked distances are given in Table 5.5 and in Figure 5.9, where, for simplicity, two variants of ten points ($N = 10$) on the plane of two attributes x_1 and x_2 are shown.

Figure 5.10 shows the procedure for sorting of clusters using the tables of interpoint distances for subsets A and B.

For each transition from one number of clusters to another, the tables of interpoint distances for each subset are rewritten such that the newly formed row in the table contains (when the poles of the dipoles are united) the shortest distance in the two cells of the preceding table. The poles of the dipoles are united in pairs for each hypothesis according to the minimum of the criterion of interpoint distance in this example.

The subsets A and B are taken into two equal parts. This is represented as an original

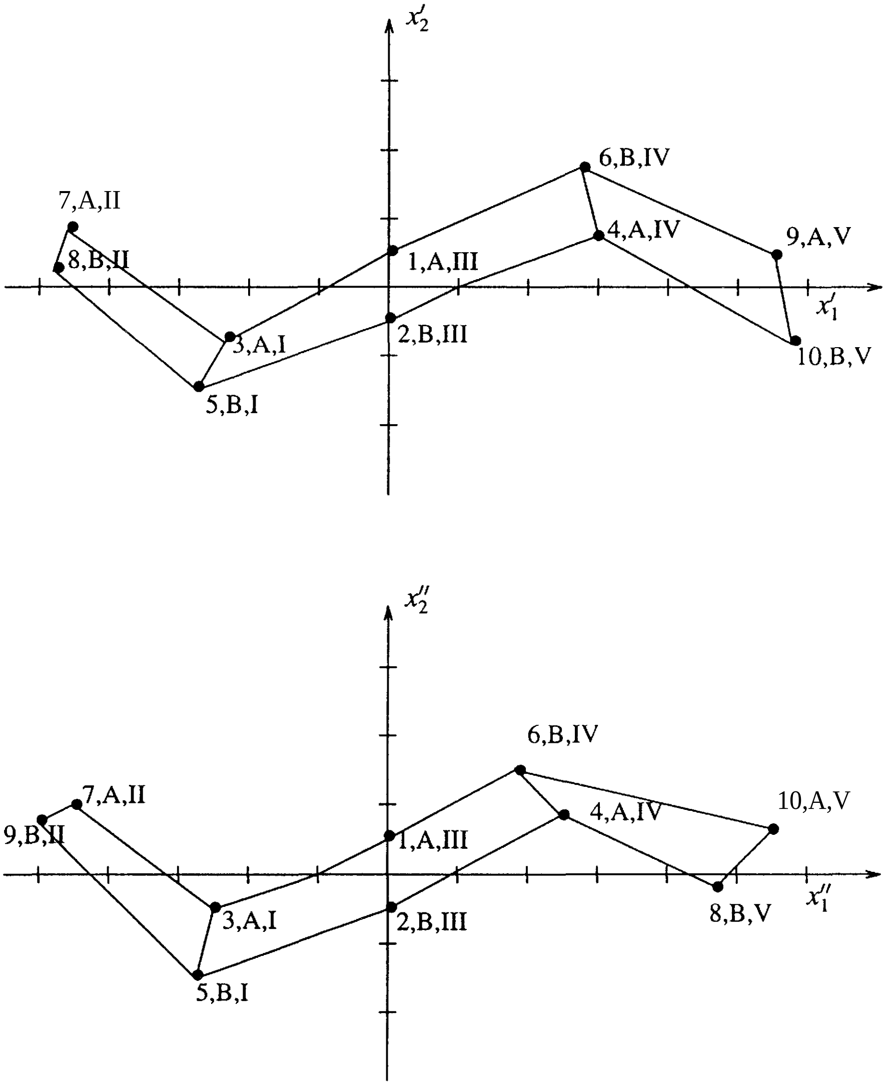


Figure 5.9. Location of the points of the two samples A, *B* in the plane; I, II, ..., V are the address of dipoles

Table 5.5. Two samples of initial data ranked by distances

No.	First sample of points			Second sample of points		
	x'_1	x'_2	$x'^2_1 + x'^2_2$	x''_1	x''_2	$x''^2_1 + x''^2_2$
1	0.00	0.40	0.16	0.00	0.40	0.16
2	0.00	-0.40	0.16	0.00	-0.40	0.16
3	-2.32	-0.69	5.86	-2.48	-0.69	6.62
4	2.80	0.68	8.30	2.54	0.785	7.07
5	-2.70	-1.25	8.85	-2.76	-1.32	9.36
6	2.60	1.60	9.32	2.52	1.78	9.52
7	-4.61	0.93	22.12	-4.40	0.90	20.17
8	-4.70	0.25	22.15	4.76	-0.10	22.67
9	5.50	0.60	30.61	-4.99	0.99	25.88
10	5.85	-0.75	34.78	5.44	0.75	30.16

code:

$$\begin{array}{cccccc}
\text{Code} & 0 & 0 & 0 & 0 & 0 \\
 & 3 & 7 & 1 & 4 & 9 \\
 & \cdot & \cdot & \cdot & \cdot & \cdot \\
 & | & | & | & | & | \\
 & \cdot & \cdot & \cdot & \cdot & \cdot \\
 & 5 & 8 & 2 & 6 & 10 \\
 & I & II & III & IV & V
\end{array}$$

(a) $k = 4$:

$$\begin{array}{cc|ccccc}
 & & I & II & III & IV & V \\
 & & 3 & 7 & 1 & 4 & 9 \\
I & 3 & 0 & 7.9 & 0 & 21 & 63 \\
II & 7 & & 0 & 22 & 55 & 102 \\
III & 1 & & & 0 & 8 & 30 \\
IV & 4 & & & & 0 & 7 \\
V & 9 & & & & & 0
\end{array}
\qquad
\begin{array}{cc|ccccc}
 & & I & II & III & IV & V \\
 & & 5 & 8 & 2 & 6 & 10 \\
I & 5 & 0 & 2 & 6 & 36 & 73 \\
II & 8 & & 0 & 22 & 51 & 112 \\
III & 2 & & & 0 & 11 & 34 \\
IV & 6 & & & & 0 & 16 \\
V & 10 & & & & & 0
\end{array}$$

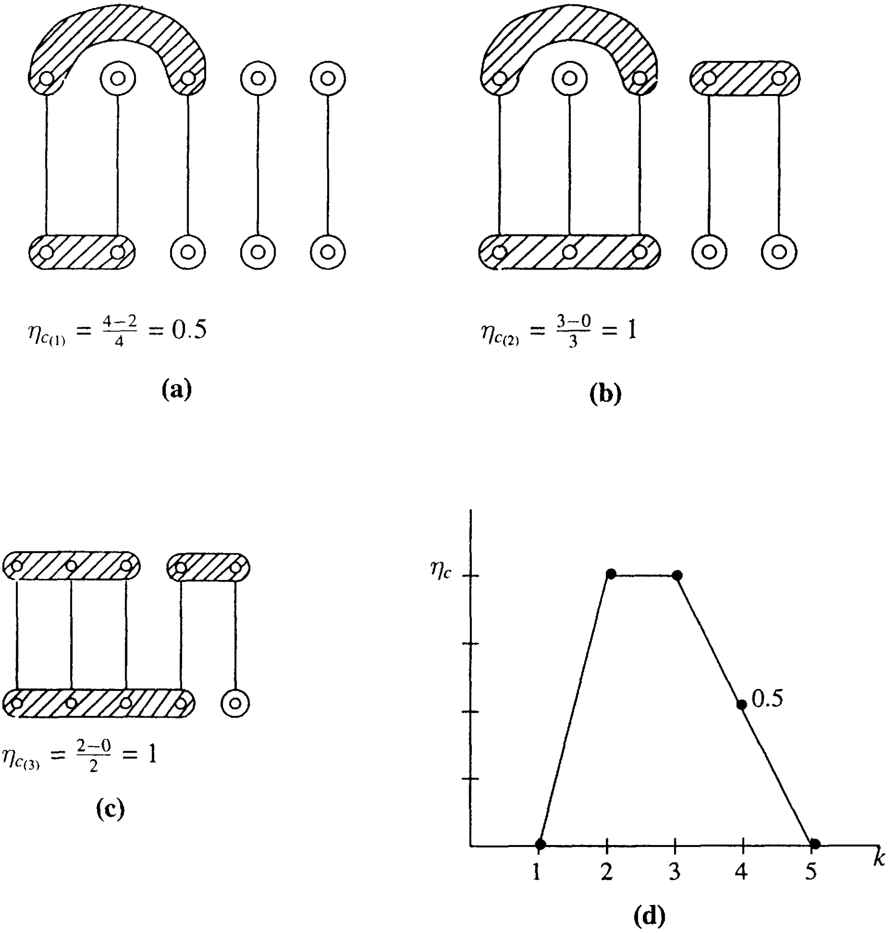
(b) $k = 3$:

$$\begin{array}{c|cccc}
 & I,III & II & IV & V \\
I,III & 0 & 7.9 & 8 & 30 \\
II & & 0 & 55 & 102 \\
IV & & & 0 & 7 \\
V & & & & 0
\end{array}
\qquad
\begin{array}{c|cccc}
 & I,II & III & IV & V \\
I,II & 0 & 6 & 36 & 73 \\
III & & 0 & 11 & 34 \\
IV & & & 0 & 16 \\
V & & & & 0
\end{array}$$

(c) $k = 2$:

$$\begin{array}{c|ccc}
 & I,III & II & IV,V \\
I,III & 0 & 7.9 & 8 \\
II & & 0 & 55 \\
IV,V & & & 0
\end{array}
\qquad
\begin{array}{c|ccc}
 & I,III & IV & V \\
I,III & 0 & 11 & 34 \\
IV & & 0 & 16 \\
V & & & 0.
\end{array}$$

It is known that the consistency criterion indicates the false consistent clusterings with the actual consistent clusterings. The false consistent clusterings; i.e., false zeros of the


$\eta_{c(1)} = \frac{4-2}{4} = 0.5$
$\eta_{c(2)} = \frac{3-0}{3} = 1$
$\eta_{c(3)} = \frac{2-0}{2} = 1$
η_c
0.5
1
2
3
4
5
k

(a)

(b)

(c)

(d)

Figure 5.10. Calculation of consistency criterion on the two equal parts of the data sample

criterion can be removed by (i) a special experimental design, the purpose of which is to form a data sample for which the criterion does not indicate false zeros and (ii) using the overall consistent criterion, which is equal to the sum of partial criteria obtained for different compositions of subsets A and B.

To sort among the hypotheses, the notations are introduced for the original data sample and to the subsets (vertex numbers) as below:

$$\begin{array}{lcccccl}
\text{Code} & 0 & 0 & 0 & 0 & 0 & \\
 & 3 & 7 & 1 & 4 & 9 & \text{subset}A \\
 & \cdot & \cdot & \cdot & \cdot & \cdot & \\
\text{Dipoles} & | & | & | & | & | & \\
 & \cdot & \cdot & \cdot & \cdot & \cdot & \\
 & 5 & 8 & 2 & 6 & 10 & \text{subset}B \\
 & I & II & III & IV & V &
\end{array}$$

where / — *V* are the dipole addresses and 00000 is the initial code for the sample. A dipole is a two-point subsample. Selected dipoles have the shortest dimension of all the feasible points of the considered sample. The code changes if the corresponding dipole changes the pole addresses in the subsets. For example,

$$\begin{array}{lccccccl}
\text{Code} & 0 & 1 & 1 & 0 & 0 & & \\
 & 3 & 8 & 2 & 4 & 9 & & \text{subset}A \\
 & \cdot & \cdot & \cdot & \cdot & \cdot & & \\
\text{Dipoles} & | & | & | & | & | & & \\
 & \cdot & \cdot & \cdot & \cdot & \cdot & & \\
 & 5 & 7 & 1 & 6 & 10 & & \text{subset}B \\
 & I & II & III & IV & V & &
\end{array}$$

The partial consistency criteria are calculated for all the variants of subset composition, and their dependencies on the number of clusters are constructed. As shown in Figure 5.11, some partitioning variants for the first sample of data points do indeed yield false zeros. This gives rise to the problem of removing false zeros of the false clusterings. Repetition of the experiment with the second sample of the data points showed that none of the 16 characteristics yields false zeros.

In this example, the consistency criterion for the selected original data sample is unimodal. One can see from Figure 5.9 that a very small variation in the locations of the sample points disturbs the unimodality. So, the above experimental design aimed at attaining criterion unimodality may lead to the required result, although it is still very sensitive. This means that a small deviation in the data leads to the formation of false value of the criterion.

Overall consistency criterion

The overall consistency criterion is the sum of the values of the partial criteria obtained for all possible compositions of subsets *A* and *B*.

$$\sum \eta = \frac{1}{L}(\eta_{c_{(1)}} + \eta_{c_{(2)}} + \cdots + \eta_{c_{(L)}}), \tag{5.9}$$

where $L = 2^{k-1}$.

Figure 5.11 demonstrates the performance of the overall consistency criterion, which does not lead to the formation of false zeros for various numbers of clusters. The experiment explains the physical meaning of the stability of the overall criterion and substantiates the basic conclusions of the coding theory as follows:

- if the overall criterion does not lead to the formation of complete zeros, then among the partial codes there is at least one that ensures the same result;
- if at least one of the codes does not form false zeros, then the overall code will also be effective; and
- for a complete sorting of the codes, one necessarily finds a partitioning into parts that leads to false zeros (the unsuccessful partitioning).

Apparently, one can apply the optimal coding theory, developed in the communication theory, for determining the optimal partitioning of a data sample into subsamples.

The goal of the experimental design is to attain the global minimum among the models. The high sensitivity to small variations in the input data and absence of unimodality

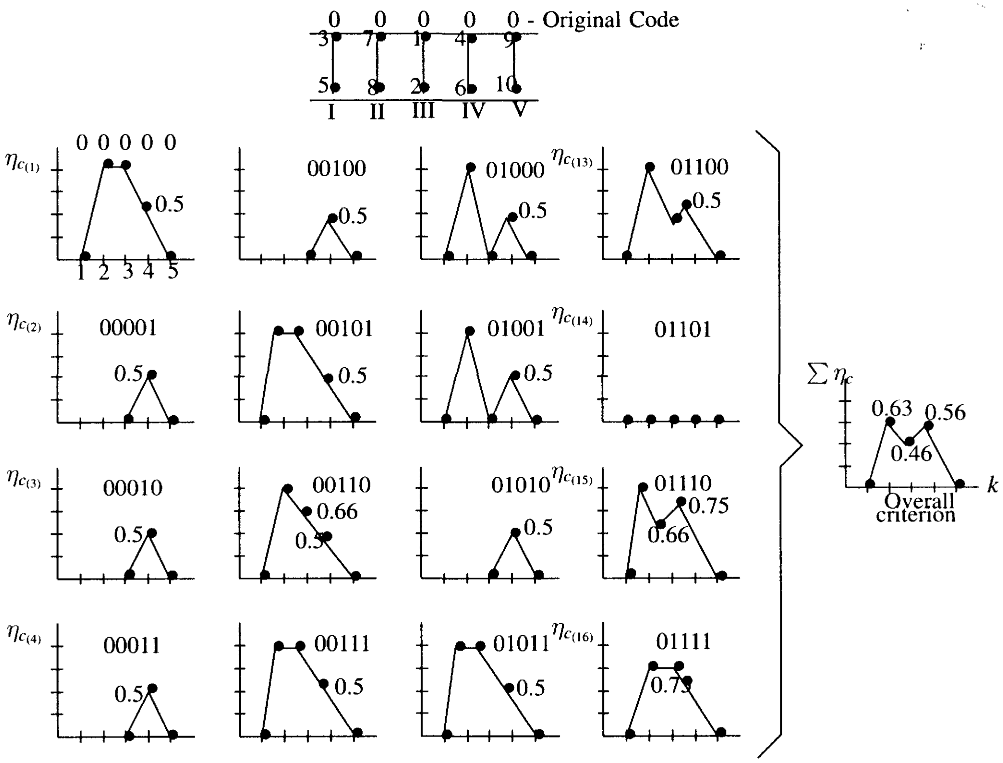


Figure 5.11. Dependence of the criterion on the number of clusters for various compositions of subsamples A and B

are characteristic symptoms of the noncorrectness of the problem of selecting a model or clusterization on the basis of a single consistency criterion. The transition to an overall consistency criterion can be viewed as one possible regularization method. With a robust approach as demonstrated above, the main goal must be the attainment of the unimodality of the consistency criterion. Sometimes, the use of the overall criterion might be insufficient in removing all the composite zeros, even for all possible partitions of the data sample into two subsets. This can be avoided by further splitting the data into subsets.

The third method of attaining unimodality consists of superimposing an auxiliary normal noise to the data sample. Its variance is increased until the most noise-immune consistent clusterization as the "locus of the minima" is achieved. One can obtain consistent clusterization without extending the experiment for regularization by the precision criterion or by experts.

Further development of this method is done by appling the canonical form of the external criterion. The locus of the minimum of the criterion coincides with the coordinates of the optimal design of the experiments and the optimal model structure. The Shannon-bias as displacement of the criterion becomes zero for all the designs and structures. This leads to a new dimension of research which will be discussed in detail in our future works.

3 OBJECTIVE COMPUTER CLUSTERING ALGORITHM

The objective computer clustering (OCC) algorithm in a generalized form is given here. The algorithm consists of the following blocks.

Block 1. Normalization of variables

Normalization is done here for the input variables $\tilde{x}_1, \tilde{x}_2, \cdots, \tilde{x}_m$, measured at N time instances as

$$x_{1i} = \frac{\tilde{x}_{1i} - \bar{x}_1}{\tilde{x}_{1_{max}} - \tilde{x}_{1_{min}}}, \; x_{2i} = \frac{\tilde{x}_{2i} - \bar{x}_2}{\tilde{x}_{2_{max}} - \tilde{x}_{2_{min}}}, \; \cdots, x_{mi} = \frac{\tilde{x}_{mi} - \bar{x}_m}{\tilde{x}_{m_{max}} - \tilde{x}_{m_{min}}}, \tag{5.10}$$

where $\bar{x}_j$, $j = 1, 2, \bullet \bullet \cdot, m$ are the mean values of corresponding variables; x_{ji}, $j = 1, 2, \cdots, m$; $i = 1, 2, \cdots, N$ are the normalized values. This can be done not only from the mean value but also from a trend of the variable. It is also useful to extend the table of attributes with the additional generalized attributes such as

$$\tilde{x}_{ij} = \frac{1}{2}(\tilde{x}_i + \tilde{x}_j), \text{ or } \tilde{x}_{ij} = \sqrt{(\tilde{x}_i\tilde{x}_j)}, \text{ or } \tilde{x}_{ij} = \sqrt{[\frac{1}{2}(\tilde{x}_i^2 + \tilde{x}_j^2)]}, \tag{5.11}$$

where $i = 1, 2, \cdots, m$; $j = i + 1, i + 2, \cdots, m$.

In addition to the input attributes, information about the goal function can be included into the original data in the form of columns with the deviated data of the output variables $y_1, y_2, \cdots, y_l$, where $l \leq m \leq m1$; and ml is the total number of primary and generalized attributes. The information about the goal function is very useful for reducing the amount of cluster search. In many clustering problems the dimension of the space / of the goal function is known: $l =$ constant. If it is not specified, it can be determined by the successive test of Karhunen-Loeve projection on to an axis, a plane, a cube, etc. or by means of the component analysis.

This is justified as follows: The modeler, while compiling the table of data, knows the goal function without fully realizing it. There necessarily exists certain axes like "good-bad," "strong-weak," "much-little," etc. These correspond to the axes serving as orthogonal projection. The space of the goal function in certain cases is two-dimensional or three-dimensional. For example, clustering of atmospheric circulation, is distinguished between two axes: the "form" and "type" of circulation; the Karhunen-Loeve orthogonal projection is applied on two variables $Y(y_1, y_2)$.

Sub-block 1a: Choose dimension of goal function

The clustering target function may be expressed by a particular vector of qualities, rather than by a scalar value. In most complex clustering problems, it is necessary to derive a complete quality vector $Y(y_1, y_2, \cdots, y_l)$.

There is a sample of observations $X(x_1, x_2, \cdots, x_m)$. Experts maintain that the target function (at any rate, one of its components—the target index) may be determined from the variance formula:

$$y = \sqrt{[\frac{1}{m}\sum_{i=1}^{m}(x_i - \bar{x}_i)^2]}, \tag{5.12}$$

where $\bar{x}_i$ is the mean value of the ith attribute.

The above formula represents the Karhunen-Loeve discrete transformation in the case where m-dimensional space of factors is mapped into one average point ("center of gravity" point, if each of the constituents has an identical mass), and the target formula is represented as a single scalar value [137]. This way, more information is retained in projecting points of an m-dimensional space onto a single axis y, although it remains a scalar quantity. The y-axis is chosen in such a way that (i) it passes through the "center of gravity" of points that is the origin of the attributes x_i, and (ii) the axis direction in the m-dimensional space is such that the points have minimum moment of inertia around the y-axis.

In the same way, even more information is retained in projecting the m-dimensional measurement space onto a two-, three-, or more dimensional spaces, to the state of projecting it on itself and not loose information. To reduce the number of computations involved in these operations, one can limit the comparisons of Karhunen-Loeve transformations to the final stage at the point on the axis or on the two-dimensional plane. The target function will be two-dimensional $Y(y_1, y_2)$, which is enough for many problems. The joint space attributes correspond to the vector of $XY(x_1, x_2, \cdots, x_m, y_1, y_2)$. This might be excessive for the optimal number of dimensions of the goal space in specific practical purposes. An optimal number of measurements for the target function space is determined by comparing the versions of the best number of coordinates that leads to consistent and accurate clusters, and by positioning these closer to the number of clusters E specified by an expert.

A way of estimating the target index. An estimation method for a single dimensional axis is developed as given below. The equation for the y-axis takes the form

$$\frac{x_1}{l_1} = \frac{x_2}{l_2} = \cdots = \frac{x_m}{l_m}, \tag{5.13}$$

where $l_1, l_2, \cdots, l_m$ are the components of the unique target vector. The moment of inertia is computed using the following criterion as

$$J_{mi} = \sum_{i=1}^{N} \sum_{j=1}^{m} x_{ij}^2 - \Big(\sum_{i=1}^{N} (l_j x_{ij})^2 / \sum_{j=1}^{m} l_j^2\Big) \to \min, \tag{5.14}$$

which amounts to the selection of $l_1, l_2, \cdots, l_m$. The second term in the criterion J_{mi} is maximal as $\sum_{i=1}^{N} \sum_{j=1}^{m} l_j x_{ji} \to$ max, with the constraints $\sum_{j=1}^{m} l_j^2 = 1$. The parameters $l_1, l_2, \cdots, l_m$ are found iteratively using the initial approximation of $l_1 = l_2 = \cdots = l_m = 1/\sqrt{m}$. This gives an equation for the y-axis. The projection of data points on the y-axis are then found. The hyperplane passing through the ith point perpendicular to the y-axis takes the orthogonal form

$$\sum_{j=1}^{m} l_j (x_j - x_{ij}) = 0, \quad i = 1, 2, \cdots, N. \tag{5.15}$$

The coordinates for the projection x_{ij_y} are determined while solving the above equation along with the equation for the y-axis. The function for allocating the projections along the i-axis is found as

$$y_i = \sqrt{}[\sum_{j=1}^{m} (x_{ij_y} - x_i)^2], \quad i = 1, 2, \cdots, N. \tag{5.16}$$

This is considered a target function and recorded in the input data.

For example, the input data corresponding to the nodes of a three-dimensional cube are shown in the Figure 5.12. The minimum value of the criterion J_{mi} corresponds to the maximum value of the function

$$(l_1 x_{11} + l_2 x_{21} + l_3 x_{31}) + (l_1 x_{12} + l_2 x_{22} + l_3 x_{32}) + \cdots + (l_1 x_{18} + l_2 x_{28} + l_3 x_{38}) \to \max . \tag{5.17}$$

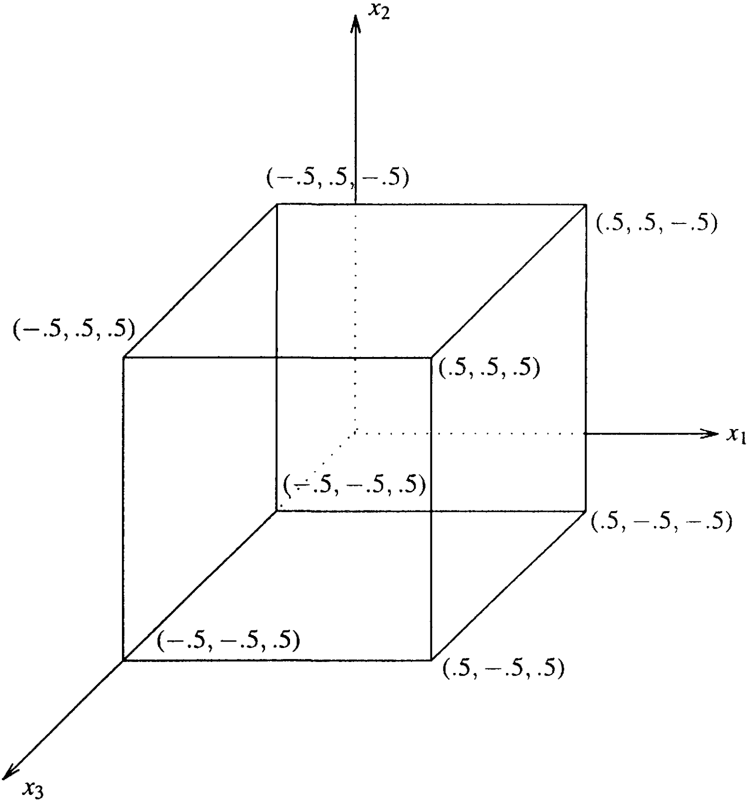


Figure 5.12. data for the given example

By iteration, $l_1 = 1, l_2 = 1$, and $l_3 = 1$ are found. The equation of the y-axis is $x_1 = x_2 = x_3$. Projections are allocated along the y-axis; at point 1, $y = +\sqrt{3/2}$, at point 8, $y = —\sqrt{3/2}$; at points 2, 3, and 4, $y = +\sqrt{3/6}$. At points 5, 6, and 7, y = $—\sqrt{3/6}$. Here, it is better not to use the Karhunen-Loeve transformation on the axis of the plane because of overlappings of many point projections. Only two projections coincide on the plane. This is solved in a different way in [124].

There is much in common between the successive application of Karhunen-Loeve projection and the method of principal components of factor analysis. The variance decreases continuously as components are isolated. Specifing a threshold is required for choosing the number of components. According to Shannon's second-limit theorem, there exists an optimal number of factors which are to be isolated. In self-organization clustering, the consistent criterion is recommended to select the optimal number of principal components; consequently, the dimension of the goal function $Y(y_1, y_2, \bullet \bullet \bullet, y_l)$ is determined.

Block 2. Calculation of variances and covariances

The data sample is given in matrix form as $X = [x_{ij}]$, $Y = [y_i]$; $i = 1, 2, \cdots, N$, $j = 1, 2, \bullet \bullet \cdot, m$. The matrix of variances and covariances G $= \frac{1}{N} X^T X$has the elements

$$g_{ij} = \text{cov}\,(x_i, x_j) = \frac{1}{N}\sum_{k=1}^{N} x_{ki}x_{kj}, \tag{5.18}$$

where x_i and x_j are the columns i and j of the matrix X.

Block 3. Isolation of effective ensembles

This is done in one of the following three ways:

Sub-block 3a. Full search over all attribute ensembles

This refers to clustering without goal function. A full search of all possible clusterings differing by the contents of the set is to be carried out in the absence of the numerical data on the goal function. For each value of the number of clusters k, 2^{m1} clusterings are to be tested using the consistency criterion, where ml is the number of attributes—including the paired or generalized attributes. This type of cluster analysis is feasible for a small number of attributes of up to $m1 = 6$. In a larger dimension of the attribute space, effective attribute ensembles are selected using the inductive learning algorithms or correlation analysis. At the same time, the goal function (scalar or vector form) must be determined experimentally by orthogonal projection. This means that it leads to clustering with goal function.

Sub-block 3b. Selection by inductive learning algorithms

This is done by using the inductive learning algorithms. The consistency criterion is used in selecting the effective attribute ensembles. The models are of the form:

$$\begin{aligned}
y_{11} &= f_{11}(x_1x_2\cdots x_{m1}), y_{21} = f_{21}(x_1x_2\cdots x_{m1}), \;\cdots,\; y_{l1} = f_{l1}(x_1x_2\cdots x_{m1}),\\
y_{12} &= f_{12}(x_1x_2\cdots x_{m1}), y_{22} = f_{22}(x_1x_2\cdots x_{m1}), \;\cdots,\; y_{l2} = f_{l2}(x_1x_2\cdots x_{m1}),\\
y_{1F} &= f_{1F}(x_1x_2\cdots x_{m1}), y_{2F} = f_{2F}(x_1x_2\cdots x_{m1}), \;\cdots,\; y_{lF} = f_{lF}(x_1x_2\cdots x_{m1}),
\end{aligned} \tag{5.19}$$

where F denotes the quantity of "freedom-of-choice." It is the number of models selected on the last layer. This indicates an ensemble of attributes for which we have to seek the most consistent clustering.

Sub-block 3c. Selection by correlation algorithm

If there are many attributes *(m* is large) and the number of measurements are small *($N < 2m$),* then it is better to use the correlation algorithm (also called "Wroslaw taxonomy") instead of inductive learning algorithms. Initially, a table of correlation coefficients of paired attributes (G) is set up. Using this matrix, the graphs of interrelated attributes for different limit values of the correlation coefficient are set up. One attribute that is correlated least with the output quantity is chosen from each graph. Ultimately, an ensemble of attributes which are correlated as little as possible with the output are determined. The limit of the correlation coefficient is gradually reduced commencing from $r_{xx} = 1$ until all attributes fall into a single path; i.e., until an ensemble containing a single attribute $y = f(x_i)$ is obtained. This way, discriminant functions which indicate effective ensembles of attributes are found:

$$y_1 = f_1(x_1x_2 \cdots x_{m1}), y_2 = f_2(x_1x_2 \cdots x_{m1}), \cdots, y_l = f_l(x_1x_2 \cdots x_{m1}).$$

Block 4. Division of data points

The ensembles obtained for different values of the correlation coefficient are subjected to a search for consistent clusterings. All ensembles are processed using the same search algorithm [124]. A square table of distances between points (with a zero diagonal) corresponding to the attributes is set up. Segments connecting any two points in the attribute space is called dipoles. These are arranged according to their length to form a full series of dipoles.

The next step is to select dipoles whose nodes form the subsets $A — B$, and $C — D$. The two nodes of the shortest dipole go into A and B; the next in magnitude go into C and D, and so on, until all nodes are investigated. Alternatively, first dipoles are chosen for A and B, and the remaining dipoles are chosen for C and D.

Half of the nodes of the dipoles go into A, while the other half go into B; subsets C and D are simply different division of the same full set of points. Conventionally, the nodes of dipoles located nearer to the coordinate origin are introduced into A and C, while those more remote are into B and D.

Block 5. Search for clusterings by consistency criterion

The next step is to carry out a search for all clusterings on the subsets A and B. Nodes belonging to the same dipole are considered equivalent. Commencing from the division of subsets into $N/2$ clusters, the number of clusters decreases to unity. The subsequent clusterings are formed by uniting into a single cluster of two points located closest to one another. The consistency criterion is determined for all clusterings by $\eta_c = (p—\Delta k)/p$, where p is the number of clusters or the number of individual points subject to clusterization, and Δk is the number of identical clusters in the subsets A and B. As a result, all clusterings for which $\eta_c = 0$ are identified. The search is repeated for all possible attribute ensembles and a map is obtained, in which consistent clusterings are denoted by dots (for example, Figure 5.8).

Additional analysis and exclusion of clusters with single dipoles. The clusters containing more than two points and the clusters containing two points belonging to the same dipole are obtained from the search of consistency criterion. The latter ones are better assigned to other clusters, or excluded from the analysis because they can represent long dipoles. Such clusters containing a single dipole are located at the end of the series of the dipoles ordered according to their length.

If the initial data table is sufficiently large (for example, $N \geq 100$, in order to avoid formation of two-points clusters), it is sufficient to use $N/3$ points instead of $N/2$ points and leave the rest of them for examining the clustering results.

Block 6. Regularization

The search is repeated on subsets C and D for further confirmation. Only those clusterings that are consistent both on A and B and on C and D are in fact considered. If we again find not one but several of the consistent clusterings, then the clustering closest to the clustering recommended by the experts is chosen. Usually, the clustering recommended by the experts turns out to be contradictory.

Block 7. Formation of output data table

The output data table that contains the division of the points of the original table into an optimal number of clusters is formed.

Block 8. Recognition

At this step, assignment of new points (images) to some cluster with the indication of the value of the goal function is carried out according to the "nearest neighbor" rule. This means that this is based on the minimum distance from the image to a point belonging to a set indicated in the initial data table.

Here we can say that the two-stage algorithm in image recognition is established in the OCC algorithms. At the first stage (teaching) of $y_i = f(x_1 x_2 \cdots x_{m1})$, the data about the space of measurements (attributes) and about the space of the goal function is used to obtain the discriminant functions with the objective of dividing the space into clusters. At the second stage (recognition), new points are assigned to some class or cluster. The number of clusters and the attribute ensemble are identified objectively using a variant search according to the consistent criterion. All the blocks given above form a schematic flow of the OCC algorithm.

Calculation of membership function of a new image to some cluster. A membership function (taken from the theory of fuzzy sets of Zadey) is given as

$$z = \frac{d_{x,i}^{-1}}{d_{x,1}^{-1} + d_{x,2}^{-1} + \cdots + d_{x,k}^{-1}} .100\%, \tag{5.20}$$

where $d_{x,i}$ is the distance from the image to the center of the cluster x; $d_{x,j}, j = 1, 2, \cdots, k$ are the distances to the centers of all clusters measured; k is the number of clusters.

The greater the membership function of an image to a cluster, the smaller is the distance from the image to the center of the cluster. The measurement of distances is carried out in the space of an effective attribute ensemble.

Example 5. Application of OCC algorithm.

The objective clustering of the rolling conditions of steel strip is considered. The original variables (x_1, x_2, x_3, x_4, and x_5) and the goal function (strip length, y) are given. It is expanded to other sets of generalized paired variables (x_6— x_{15}).

Block 1. Table 5.6 has been obtained as a result of normalization of the variables as deviations from their mean values.

Block 2. The matrix of variances and covariances is given in Table 5.7.

Block 3c Isolation of the effective attribute ensembles by the correlation algorithm of "Wroslaw taxonomy" yielded the 15 effective ensembles shown in Table 5.8.

Block 4 Division of the data according to the dipole search for the ensemble $x_5 x_{11} x_{12} x_{13}$ is as follows:

subset A: 12, 23, 38, 37, 14, 27, 15, 24, 39, 19, 28, 11, 16, 29, 20, 34, 3, 22, 25, 40;
subset B: 13, 18, 31, 32, 8, 26, 10, 17, 35, 4, 30, 7, 21, 33, 9, 36, 5, 1, 6, 2;
subset C: 32, 14, 23, 21, 38, 24, 16, 31, 22, 13, 17, 8, 34, 28, 26, 20, 18, 10, 33, 7;
subset D: 36, 11, 25, 12, 39, 15, 27, 35, 9, 4, 19, 3, 37, 40, 30, 1, 6, 5, 2, 29.

Block 5. The cluster search is carried out using the consistency criterion by dividing the subsets into eight.

Table 5.6. Normalized initial data

No.	x_1	x_2	x_3	x_4	x_5	x_6 – x_{13}	x_{14}	x_{15}	y
1	0.286	0.374	0.302	0.303	-.137	...	-.077	-.081	0.338
2	-.055	-.091	0.322	0.619	0.706	...	0.766	0.771	0.111
3	0.098	-.019	0.222	0.249	-.043	...	0.003	0.001	0.026
4	0.125	0.131	0.202	0.216	-.075	...	-.033	-.035	0.093
5	0.034	-.077	0.675	0.195	0.097	...	0.236	0.127	-.127
.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.
37	-.144	-.217	-.300	-.343	-.106	...	-.166	-.157	-.110
38	0.000	0.016	-.275	-.280	-.026	...	-.082	-.073	-.005
39	-.117	-.040	-.276	-.205	-.043	...	-.098	-.089	-.048
40	0.250	0.346	-.200	-.205	0.004	...	-.054	-.044	0.381

Table 5.7. Matrix of variances and paired variances

Attributes	x\	x_2	x_3	x_4	x_5	...	x_{14}	X\5	y
x\	0.0518	0.0525	0.0086	0.0054	-.0078	...	-.0059	-.0064	0.0515
X2		0.0602	0.0051	0.0022	-.0077	...	-.0066	-.0070	0.0553
x_3			0.0492	0.0433	-.0028	...	0.0072	0.0045	0.0047
x_4				0.0519	0.0114	...	0.0200	0.0194	0.0044
x_5					0.0481	...	0.0474	0.0474	-.0058
.						.	.	.	.
.						.	.	.	.
.						.	.	.	.
x_{14}							0.0482	0.0479	-.0049
x_{15}								0.0480	-.0048
y									0.0569

Table 5.8. Effective attribute ensembles

No.	Ensembler
1	x_{11}
2	$x_5\ x_{11}$
3	$x_5\ x_{11}\ x_{13}$
4	$x_5\ x_{11}\ x_{12}\ x_{13}$
5	$x_5\ x_9\ x_{11}\ x_{12}\ x_{13}$
6	$x_5\ x_9\ x_{11}\ x_{12}\ x_{13}\ x_{14}$
7	$x_5\ x_9\ x_{11}\ x_{12}\ x_{13}\ x_{14}\ x_{15}$
8	$x_1\ x_5\ x_9\ x_{11}\ x_{12}\ x_{13}\ x_{14}\ x_{15}$
9	$x_1\ x_5\ x_8\ x_9\ x_{11}\ x_{12}\ x_{13}\ x_{14}\ x_{15}$
10	$x_1\ x_5\ x_7\ x_8\ x_9\ x_{11}\ x_{12}\ x_{13}\ x_{14}\ x_{15}$
11	$x_1\ x_5\ x_6\ x_7\ x_8\ x_9\ x_{11}\ x_{12}\ x_{13}\ x_{14}\ x_{15}$
12	$x_1\ x_3\ x_5\ x_6\ x_7\ x_8\ x_9\ x_{11}\ x_{12}\ x_{13}\ x_{14}\ x_{15}$
13	$x_1\ x_3\ x_4\ x_5\ x_6\ x_7\ x_8\ x_9\ x_{11}\ x_{12}\ x_{13}\ x_{14}\ x_{15}$
14	$x_1\ x_2\ x_3\ x_4\ x_5\ x_6\ x_7\ x_8\ x_9\ x_{11}\ x_{12}\ x_{13}\ x_{14}\ x_{15}$
15	$x_1\ x_2\ x_3\ x_4\ x_5\ x_6\ x_7\ x_8\ x_9\ x_{10}\ x_{11}\ x_{12}\ x_{13}\ x_{14}\ x_{15}$

Block 6. The consistent clusters are further determined by the condition of their presence on the maps obtained for the subsets A, B and C, D and summarized on the summary map as shown in Figure 5.13. The clustering marked C in the figure is the most effective one.

Block 7. The following data points are grouped into clusters according to the mean strip length by using the above result of objective clustering.

Cluster 1: Points 6, 18, 23, and 25 for y = 10.99;
Cluster 2: Points 2, 29, and 33 for y = 11.63; and
Cluster 3: Points 1, 3, 4, 5, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 19, 20, 21, 22, 24, 26, 27, 28, 30, 31, 32, 34, 35, 36, 37, 38, 39, and 40 for y = 11.77.

Block 8. In the recognition stage, let us assume that a new image is obtained with the attribute values of x_5 = 4.373, x_{11} = 26.986, x_{12} = 6.631, and x_{13} = 70.202. Then the distances from the point obtained to all 40 initial points are calculated. The nearest point is located as the point 30 with the attribute values of x_5 = 4.410, x_{11} = 26.96, x_{12} = 6.65, and x_{13} = 70.28. This point belongs to the third cluster; consequently, the new point image belongs to the third cluster. The values of the membership function reveal that the first cluster z - 0.203, the second cluster z = 0.240, and the third cluster z = 0.553; i.e., the input image affiliates more to the third cluster.

4 LEVELS OF DISCRETIZATION AND BALANCE CRITERION

The criteria of differential type are quite varied, but they, nonetheless, ensure the basic requirement of Gödel's approach. They are a clustering found by sorting according to a criterion using a new data set which is not used with the internal criterion. In the algorithms described above, the basic criterion used is consistency. Here is another form of differential criterion: the criterion of balance of discretization is proposed for selecting optimal clusterings in self-organization clustering algorithms for a varying degree of fuzziness of the mathematical description language [34]. The principle behind this criterion is that the overall picture of the arrangement of the clusters in the multidimensional space of features must not differ greatly from the type of discretization of the variable attributes. The optimal clustering (the number of clusters and the set of features) must be the same—independent of the number of levels of discretization of the variables indicated in the data sample.

Initial data sample is discretized into various levels on the coordinate axes to find the optimal clustering. Hierarchical trees for sorting the number of clusters are set up from the tables of interpoint distances. The optimal number of clusters coincides at the higher levels of hierarchy of reading variables. The balance of discretization criterion is used like the criterion of consistency; i.e., according to the number of identical clusters.

In self-organization modeling the criterion of consistency, which is called the minimum-bias criterion to estimate the balance of structures, is computed according to the formula $\eta_{bs} = \sum_{i=1}^{N}(\hat{y}^A - \hat{y}^B)^2$. The criterion requires that the model obtained for the subset A $(\hat{y}^A)$ differs as little as possible from the model obtained for the subset B $(\hat{y}^B)$. If the criterion has several equal minima (balances), then we have to apply some method of regularization.

In self-organization clustering, the data sample is discretized into different numbers of levels according to the coordinates of the points for obtaining subsets A and B. It is then sorted among the hypotheses as to the number of clusters for each of the subsets and the results compared with one another. The optimal clustering corresponds to the minimum of the consistency criterion; usually its zero value resembles the balance of clusterings on both the subsets.

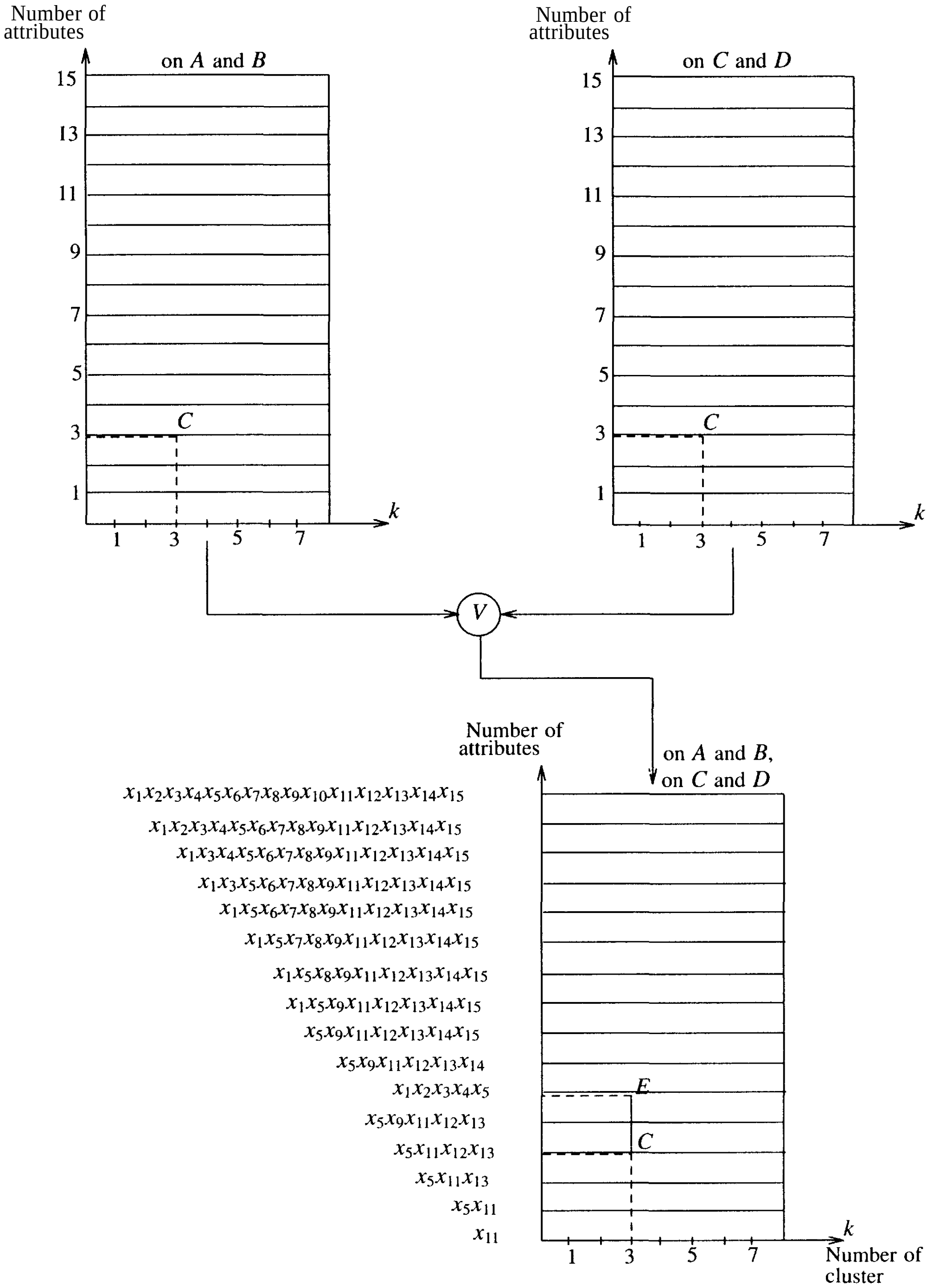

Number of attributes
on A and B
15
13
11
9
7
5
3
1
C
1 3 5 7
k
Number of attributes
on C and D
15
13
11
9
7
5
3
1
C
1 3 5 7
k
V
Number of attributes
on A and B, on C and D
$x_1x_2x_3x_4x_5x_6x_7x_8x_9x_{10}x_{11}x_{12}x_{13}x_{14}x_{15}$
$x_1x_2x_3x_4x_5x_6x_7x_8x_9x_{11}x_{12}x_{13}x_{14}x_{15}$
$x_1x_3x_4x_5x_6x_7x_8x_9x_{11}x_{12}x_{13}x_{14}x_{15}$
$x_1x_3x_5x_6x_7x_8x_9x_{11}x_{12}x_{13}x_{14}x_{15}$
$x_1x_5x_6x_7x_8x_9x_{11}x_{12}x_{13}x_{14}x_{15}$
$x_1x_5x_7x_8x_9x_{11}x_{12}x_{13}x_{14}x_{15}$
$x_1x_5x_8x_9x_{11}x_{12}x_{13}x_{14}x_{15}$
$x_1x_5x_9x_{11}x_{12}x_{13}x_{14}x_{15}$
$x_5x_9x_{11}x_{12}x_{13}x_{14}x_{15}$
$x_5x_9x_{11}x_{12}x_{13}x_{14}$
$x_1x_2x_3x_4x_5$
$x_5x_9x_{11}x_{12}x_{13}$
$x_5x_{11}x_{12}x_{13}$
$x_5x_{11}x_{13}$
x_5x_{11}
x_{11}
E
C
1 3 5 7
k
Number of cluster

Figure 5.13. Maps of location of consistent clusterizations

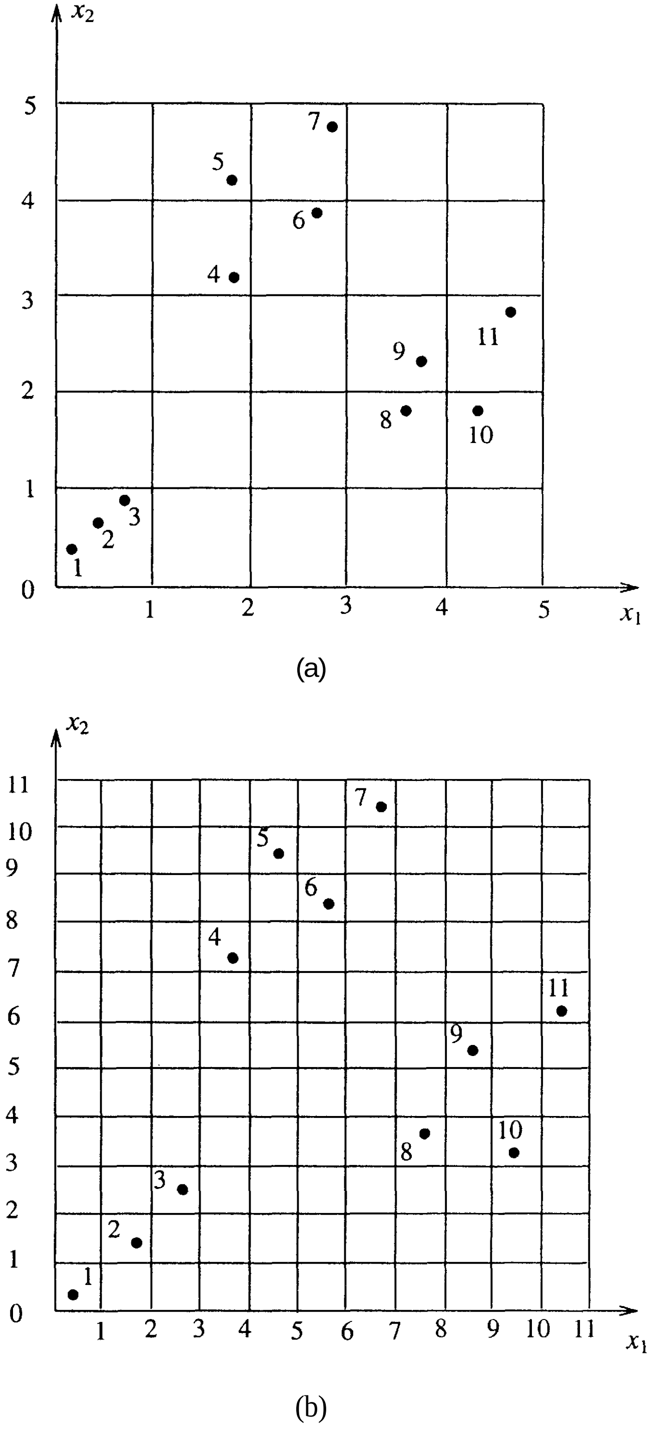


Figure 5.14. Discretization of the coordinates x_1 and x_2 at the levels of (a) five and (b) eleven

Levels of discretization

Figure 5.14 illustrates the different levels of discretization of the coordinates of the points x_1 and x_2 according to Widrow's recommendations. It is suggested that the number of discretization levels of the multiples correspond to obtaining the false zeros of the criterion; for example, here it is $N_1 = N = 11$ and $N_2 = N/2 - 1 = 5$ levels.

In computing the criterion of consistency or balance of discretizations, one has to carry out a special procedure of superimposing square matrices of interpoint distances. The following matrices are obtained according to the 11th and 5th levels of discretizations.

	1	2	3	4	5	6	7	8	9	10	11
1	0	2	4	10	13	13	16	10	13	13	16
2		0	2	8	11	11	14	8	11	11	14
3			0	6	9	9	12	6	9	9	12
4				0	3	3	6	8	7	9	8
5					0	2	3	9	8	10	9
6						0	3	7	6	8	7
7							0	8	7	9	8
8								0	3	3	6
9									0	2	3
10										0	3
11											0

	1	2	3	4	5	6	7	8	9	10	11
1	0	0	0	4	5	5	6	4	5	5	6
2		0	0	4	5	5	6	4	5	5	6
3			0	4	5	5	6	4	5	5	6
4				0	1	1	2	4	3	5	4
5					0	2	1	5	4	6	5
6						0	1	3	2	4	3
7							0	4	3	5	4
8								0	1	1	2
9									0	2	1
10										0	1
11											0

The following matrix shows the inter-cluster distances of clusters from both of the above tables. The table for five levels does not differ essentially from the table for eleven levels.

	1, 2, 3	4, 5, 6, 7	8, 9, 10, 11
1, 2, 3	0	6	6
4, 5, 6, 7		0	6
8, 9, 10, 11			0

Calculation of the criterion

The criterion of balance of discretization is calculated in a special way, which is very convenient for programming. This is done at each step of the construction of hierarchical trees *for* sorting hypotheses as to the number of clusters. *The* points that make a cfuster are marked with indices (vertices) in a space *of* N x N matrices for subsets A and B. The criterion is computed as

$$B_L = \frac{(k - \Delta k)}{k}, \tag{5.21}$$

where $k = N^2$ and Δk is the number of coincidence points or indices on the marking spaces.

The final values are trivial and always hold good. It gives $B_L = 0$ for the optimal clusters, which corresponds to our human impressions when looking at the given arrangement of points.

Regularization

If in the interval from $k = 1$ to $k = N/2$ several zero values of the criterion B_L are formed (excluding ends of the interval), it is necessary to determine which of the "zeros" are false and which are true. This can be checked by repeating the construction of the sorting tree for the hypotheses from some intermediate number of levels (for example, seven or eight if it was checked for 11 before). The whole procedure does not cause any special difficulties for larger number of points and levels.

Example 6. Optimal clustering using the criterion of balance of discretization.

The data is given in Figure 5.14b for the attributes $x\backslash$ and $X2$ at the discretization level of 11. The table of interpoint distances for the entire sample is measured as given in the matrix

$$\begin{array}{c|ccccccccccc}
 & 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9 & 10 & 11 \\
1 & 0 & 1 & 2 & 7 & 9 & 9 & 11 & 7 & 8 & 9 & 11 \\
2 & & 0 & 1 & 6 & 8 & 8 & 10 & 7 & 8 & 9 & 10 \\
3 & & & 0 & 6 & 8 & 7 & 10 & 6 & 7 & 8 & 9 \\
4 & & & & 0 & 2 & 2 & 4 & 5 & 4 & 7 & 7 \\
5 & & & & & 0 & 2 & 2 & 7 & 6 & 8 & 7 \\
6 & & & & & & 0 & 3 & 5 & 4 & 6 & 4 \\
7 & & & & & & & 0 & 7 & 6 & 7 & 5 \\
8 & & & & & & & & 0 & 2 & 2 & 3 \\
9 & & & & & & & & & 0 & 2 & 3 \\
10 & & & & & & & & & & 0 & 2 \\
11 & & & & & & & & & & & 0
\end{array}$$

The dipoles are constructed so that they start with the shortest until all the points are in the subsets A and B without repeating them. The following dipoles are obtained and formed into subsets A and B.

$$\begin{array}{lccccccccccc}
\text{subset A:} & 1 & 2 & 1 & 4 & 4 & 5 & 5 & 8 & 8 & 9 & 10 \\
 & . & . & . & . & . & . & . & . & . & . & . \\
 & | & | & | & | & | & | & | & | & | & | & | \\
 & . & . & . & . & . & . & . & . & . & . & . \\
\text{subset B:} & 2 & 3 & 3 & 5 & 6 & 6 & 7 & 9 & 10 & 10 & 11
\end{array}$$

They are addressed as $I, II, III, IV, V, VI, VII, VIII, IX, X, XI$. The matrices of interpoint distances are compiled for the subsets A and B separately as below:

		I	*II*	*III*	*IV*	*V*	*VI*	*VII*	*VIII*	*IX*	*X*	*XI*
	A	1	2	1	4	4	5	5	8	8	9	10
I	1	0	1	0	7	7	9	9	7	7	8	9
II	2		0	1	6	6	8	8	7	7	8	9
III	1			0	7	7	9	9	7	7	8	9
IV	4				0	0	2	2	5	5	4	7
V	4					0	2	2	5	5	4	7
VI	5						0	0	7	7	6	8
VII	5							0	7	7	6	8
VIII	8								0	0	2	2
IX	8									0	2	2
X	9										0	2
XI	10											0

		I	*II*	*III*	*IV*	*V*	*VI*	*VII*	*VIII*	*IX*	*X*	*XI*
	B	2	3	3	5	6	6	7	9	10	10	11
I	2	0	1	1	8	8	8	10	8	9	9	10
II	3		0	0	8	7	7	10	7	8	8	9
III	3			0	8	7	7	10	7	8	8	9
IV	5				0	2	2	2	6	8	8	7
V	6					0	0	3	4	6	6	4
VI	6						0	3	4	6	6	4
VII	7							0	4	7	7	3
VIII	9								0	2	2	2
IX	10									0	0	2
X	10										0	2
XI	11											0

Two hierarchical trees of sorting hypotheses as to the clusters (figure 5.15) are built up using the compiled interpoint distance matrices. The criterion of balance of discretization is calculated at each step of constructing the hierarchical trees. The vertices of the dipoles are combined in the tree into a cluster. The elements of the clusters are marked with indices or circles in the matrix form as mapped out in Figure 5.16. Superimposition of the matrix constructed for subset A on the matrix constructed for subset B makes it possible to compute the criterion $B_L = (k - \Delta k)/k$, where $k = N^2 = 121$ and Δk is the number of cells that are coinciding in the matrices.

The "zero" values for the criterion are found for $k_{(A)} = k_{(B)} = 1, 3, 5,$ and 11 by comparing both the trees.

If there are several "zero" values of the criterion, then one has to "invert" certain dipoles and calculate the overall criterion of consistency or one has to repeat the procedure with the different number of levels of discretization.

The examples described in this chapter show that sorting according to the differential criteria (having the properties of the external criteria), consistency, and balance of discretization can replace a human expert in arriving at subjective notions regarding the number and composition of points of the clusters.

5 FORECASTING METHODS OF ANALOGUES

In the traditional deductive methods of modeling, specifying the output and input variables is usually required. The number of variables is equal to or less than the number of data

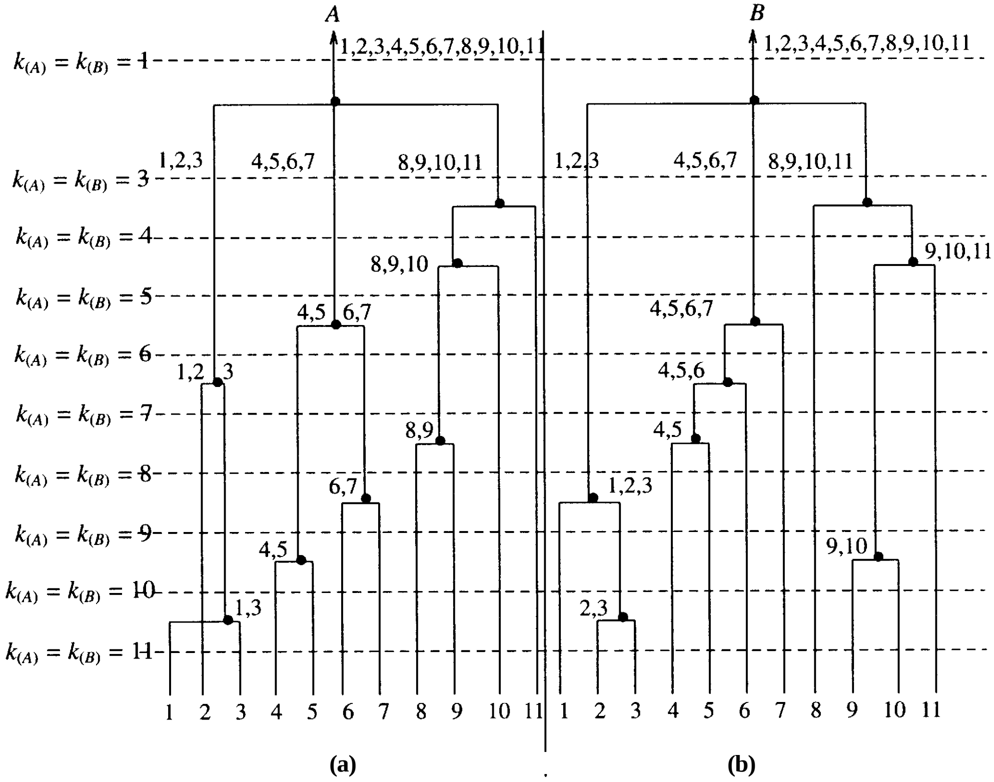
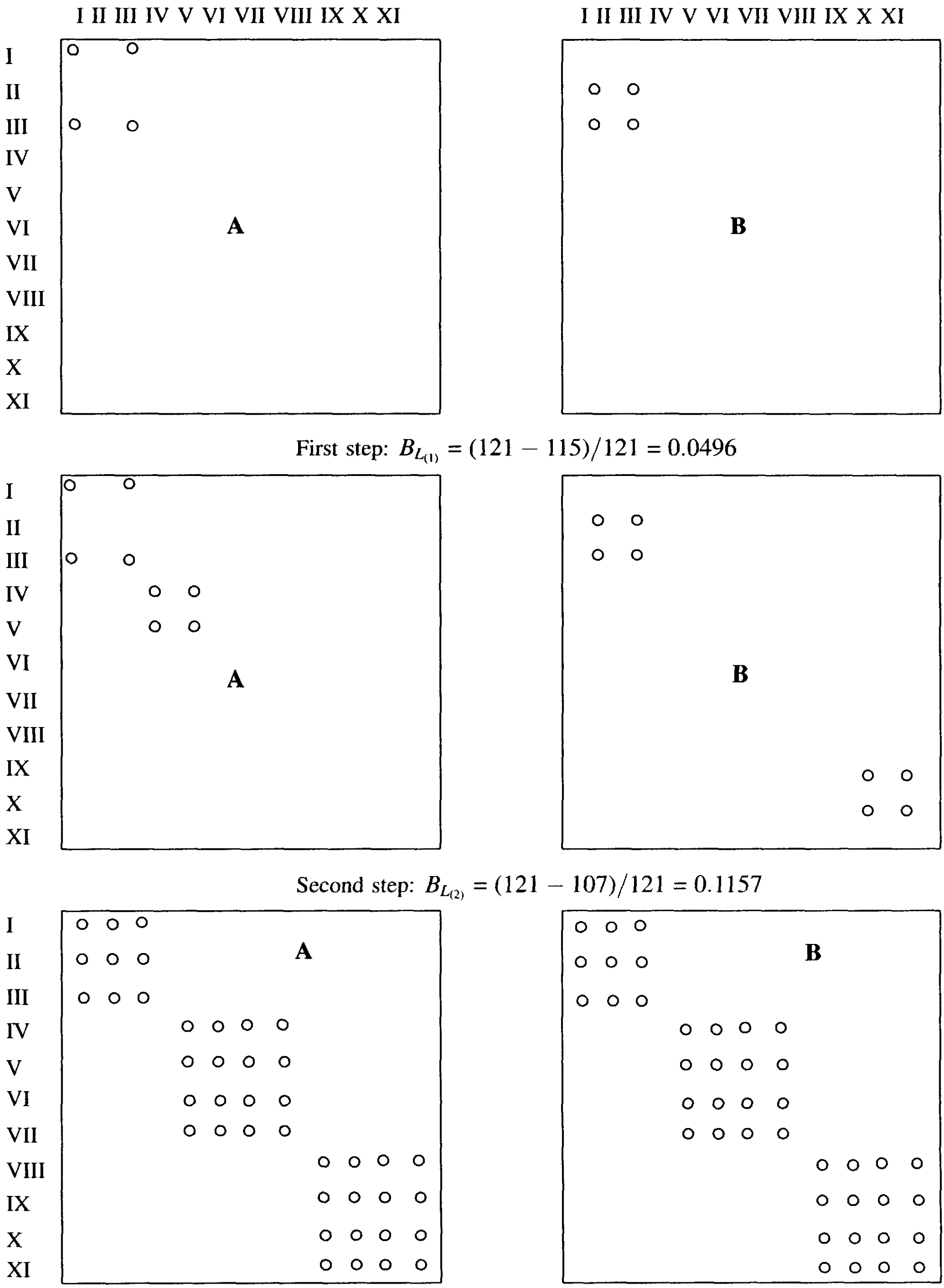


Figure 5.15. Hierarchical trees of sorting hypotheses as to the number of clusters using different discretization levels

measurements. In regression analysis, there are additional limitations, such as the noise factor affecting the output variable, the regressor set being complete, and the regressors not taking into account the equation operate as additional noise. The theories of principal component analysis and pattern analysis for predicting biological, ecological, economic, and social systems which have proven to be possible in a fuzzy language are not new. Again, this is based on the deductive principle that the more fuzzy the mathematical language of prediction, the longer its maximum achievable anticipation time.

Unlike deductive algorithms, the objective system analysis (OSA) algorithm has additional advantages. This does not require an output variable to be specified. In turn, all variables are considered as output variables and the best variant is chosen by the external criterion. The weak point of the inductive learning algorithms is that the estimate of parameters is done by means of the regression analysis. The limitations of the regression analysis cannot be overcome even by using the orthogonal polynomials. The resultant expectations of estimators are biased both by noise in the initial data and the incomplete number of input variables. A physical model is the simplest one among unbiased ones derived with the exact data or with the infinitely large data sample.

Nonparametric inductive learning algorithms offer another possibility and promise to be more effective than the deductive and parametric inductive ones. Its approach is to clarify that in the area of complex systems modeling and forecasting where objects and their mathematical models are ill-defined, the optimum results are achieved as the degree of "fuzzyness" of a model is adequate to the "fuzzyness" of an object. This means that the

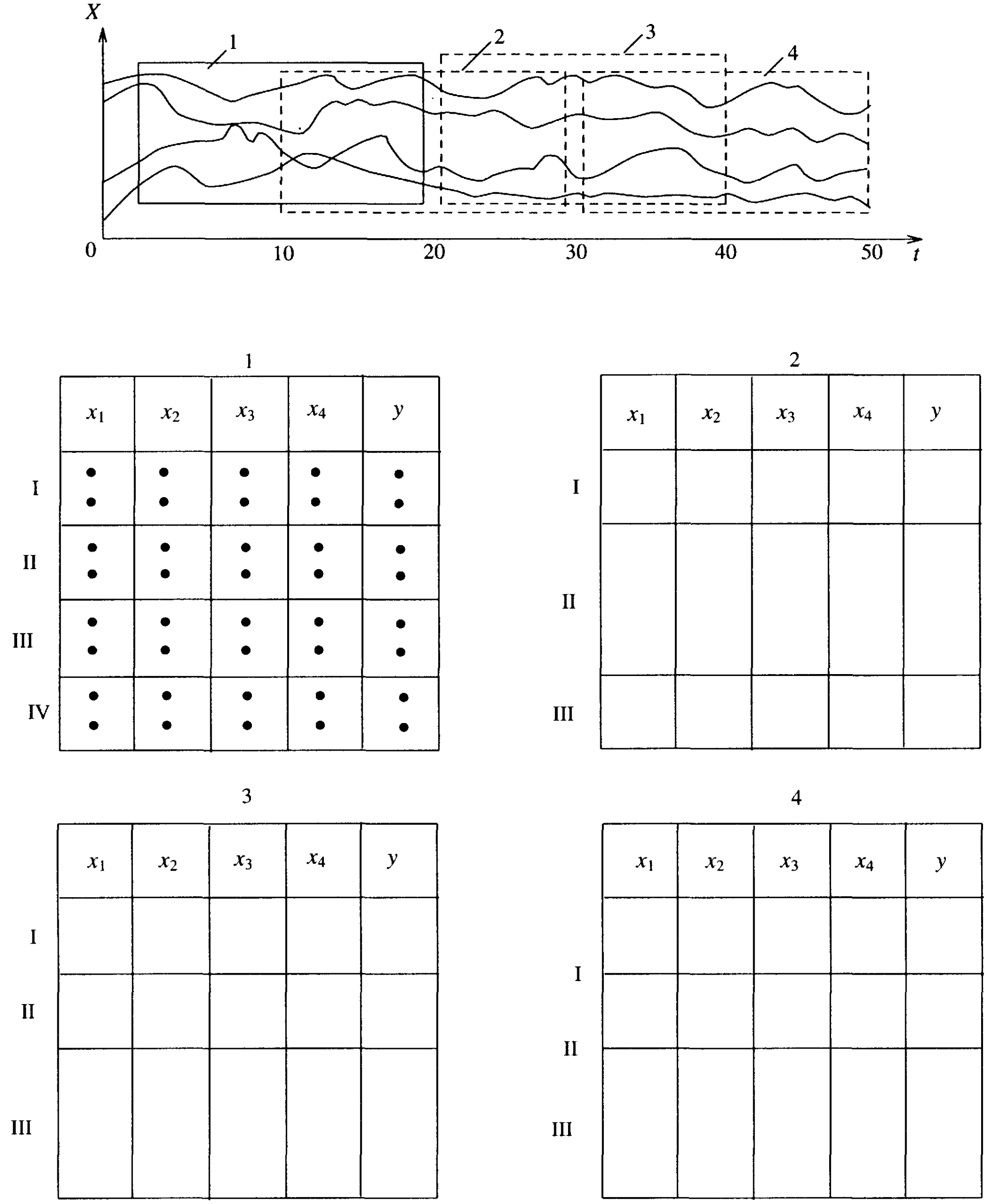

I II III IV V VI VII VIII IX X XI
I II III IV V VI VII VIII IX X XI
I
II
III
IV
V
VI
VII
VIII
IX
X
XI
A
B

First step: $B_{L_{(1)}} = (121 - 115)/121 = 0.0496$

Second step: $B_{L_{(2)}} = (121 - 107)/121 = 0.1157$

Ninth step: $B_{L_{(9)}} = (121 - 121)/121 = 0$

Figure 5.16. Calculation of the consistency criterion from the mappings



1

	x_1	x_2	x_3	x_4	y
I	● ●	● ●	● ●	● ●	● ●
II	● ●	● ●	● ●	● ●	● ●
III	● ●	● ●	● ●	● ●	● ●
IV	● ●	● ●	● ●	● ●	● ●

2

	x_1	x_2	x_3	x_4	y
I					
II					
III					

3

	x_1	x_2	x_3	x_4	y
I					
II					
III					

4

	x_1	x_2	x_3	x_4	y
I					
II					
III					

Figure 5.17. Four positions of the "sliding window" and coresponding four clusterizations (number of clusters decreases from four to three)

equal “fuzzyness” is reached automatically if the object itself is used for forecasting. This is done by searching analogues from the given data sample as the clusterizations are tracked using a "sliding window" that moves along the data sample in time axis. For example, the data sample for the ecosystem of Lake Baykal contains measurements over an interval of 50 years (Figure 5.17). One can obtain 40 clusterization forms used to track how the ecological system varies by moving a 10-year wide sliding window in order to predict its further developement. The longest anticipation time of a prediction is obtained without using any polynomial formulations. The objective clusterization of the given data sample is used to calculate the graph of the probability of transition from one class to another. This makes it possible to find an analogue of the current state of the object in prehistory and, consequently, to indicate the long-term prediction. It follows that the choice of the number of clusters is a convenient method of changing the degree of fuzziness in the mathematical language description of the object. By varying the width of the "sliding window," one can realize an analogous action in the choice of the patterns. This approach has an advantage over the clustering analysis given by the OCC algorithm and also the OSA algorithm for having a minimum number of points.

5.1 Group analogues for process forecasting

The method of group analogues leads to the solution of the forecasting problem of a multidimensional process by pattern and cluster search with a subsequent development of a weak into a detailed forecast by the forecasting method of analogues. A sample of observations (N) of a multidimensional process serves as the initial data, and the set of measured variables $(x_1, x_2, \cdots, x_m)$ is sufficiently representative; i.e., it characterizes the state of the observed object and what has occurred in the past is repeated in the present if the initial state has been analogous.

In the problems of ecology, economics, or sociology the available sample size is usually small. The number of forecast characteristic variables ra is significantly larger than the number of sample points N $(N \ll m)$. Nevertheless, the forecasts are necessary and are of the basic means of increasing their effectiveness through the use of the "method of group analogues."

Forecasts are not calculated, but selected from the table of observation data. This opens up the possibility of more successful forecasting of multidimensional processes.

Formula for forecast measure

The forecasting accuracy of each variable is characterized by the forecast variation of

$$\delta^2_{i_{(k)}} = \sum_{k=1}^{N_C} (x_{i_{(k)}} - \hat{x}_{i_{(k)}})^2 / (x_{i_{(k)}} - \bar{x}_{i_{(k)}})^2, \tag{5.22}$$

where $x_{i_{(k)}}$ is the actual value of the ith variable, $\hat{x}_{i_{(k)}}$ is the forecast obtained as explained below, and $\bar{x}_{i_{(k)}}$ is the mean value (for a quasi-stationary process) without taking the forecast point into account. If the process is nonstationary; i.e., if some of the variables have a clear expression of trend (they increase or decrease continuously), then $x_{i_{(k}}$ equals the value of the trend at each forecasting step. The above formula compares the average error of the forecast by the analogues method with respect to the average error of the forecast as the mean value or trend value.

The forecast of each variable is considered to be successful if the variation $\delta^2_{i_{(k)}} \leq 1.0$ (or in percentage, $<$ 100%). Usually, only some variables forecast well. In the best case for all m variables $\delta^2_{i_{(k)}} = 1.0$ (or = 100%). To successfully increase this percentage of forecast variables for a short sample of initial data, one has to go from a search for one analogue in prehistory to the problem of combining several analogues.

Forecast space of several analogues

Here x_i is the point in the multidimensional (Euclidean) space of variables and $\hat{x}_i$, in the space of forecasts, corresponds to each row of the table of initial data sample. The former space is used for computing the interpoint distances, while the latter is used to approximate the forecasts by splines or polynomial formulations.

The point B of the multidimensional spaces x_i and $\hat{x}_i$ is denoted as the output point for forecasting. This is either the last point of the sample in time or the last one that would be possible in estimating the variation of the obtained forecast by the last row. The distances between the point B and all other points measured in the space x_i determine the possibility of using them as analogues. The closest point A_1 is called the first analogue, the next one in distance A_2 is called second analogue, and so on until the last analogue A_F $(F \leq N)$. A specific forecast $\hat{x}_i$ corresponds in the forecast space to each analogue. The number of analogues are combined—either specified by an expert or determined according to an inductive algorithm. Various methods can be proposed. Here the method based on extrapolating the forecast space by splines is considered. It is assumed that some forecast value, which is determined by using the forecasts at adjacent points of the space, exists at each point of the forecast space $\hat{x}_i$.

"Combining" forecasts by splines

Here "combining" means approximating the data by splines or polynomial equations with a subsequent calculation of the forecast at the point B. The forecast is defined with the help of weighted summing of forecast analogues using spline equations

$$\begin{aligned}\hat{x}_{i_{(B)}} &= f(\hat{x}_{i_{(A_1)}}, \hat{x}_{i_{(A_2)}}, \cdots, \hat{x}_{i_{(A_F)}}) \\ &= a_0 + a_1\hat{x}_{i_{(A_1)}} + a_2\hat{x}_{i_{(A_2)}} + \cdots + a_F\hat{x}_{i_{(A_F)}}. \end{aligned} \tag{5.23}$$

The splines are selected such that the point B approaches the optimal set of analogues A_s $(1 < s < F)$; i.e., the difference between their forecasts decreases. The closer the points in the forecast space $\hat{x}_i$ are, the closer are the forecasts themselves at these points.

Distances between points for a short-range one-step forecast are measured in the space x_i as below:

$$d_j = \sqrt{[(x_1^{(A_j)} - x_1^{(B)})^2 + (x_2^{(A_j)} - x_2^{(B)})^2 + \cdots + (x_m^{(A_j)} - x_m^{(B)})^2]}; \qquad \text{where } j = 1, 2, 3, \cdots, F, \tag{5.24}$$

where d_j are the Euclidean distances of the point B from the analogues A_j, $j = 1, 2, \cdots, F$; A_1 is the first analogue (closest), A_2 is the second more distant analogue, A_3 is the third even more distant analogue, and so on.

The Euclidean distance is a convenient measure of proximity of a point, but only for a one-step forecast. The repetitive procedure of stepwise forecast can be used to obtain a long-range forecast with a multi-step lead, in which a "correlative measure" is estimated for the proximity of groups of points. The canonical correlation coefficient [104] is also recommended as a proximity measure for forecasting more than four steps.

The interpoint distances d_j, $j = 1, 2, \cdots, F$ are used for calculating the coefficients of the following splines;

1. for one analogue $(F = 1)$:

$$\hat{x}_{i_{(B)}} = \hat{x}_{i_{(A_1)}}; \tag{5.25}$$

2. when two analogues are taken into consideration (F = 2):

$$\hat{x}_{i_{(B)}} = (d_1^{-1}\hat{x}_{i_{(A_1)}} + d_2^{-1}\hat{x}_{i_{(A_2)}})/(d_1^{-1} + d_2^{-1}); \tag{5.26}$$

3. when the forecasts of three analogues are taken into account (F = 3):

$$\hat{x}_{i_{(B)}} = (d_1^{-1}\hat{x}_{i_{(A_1)}} + d_2^{-1}\hat{x}_{i_{(A_2)}} + d_3^{-1}\hat{x}_{i_{(A_3)}})/(d_1^{-1} + d_2^{-1} + d_3^{-1}); \tag{5.27}$$

4. when the forecasts of F analogues are taken into account:

$$\hat{x}_{i_{(B)}} = (\sum_{\alpha=1}^{F} d_\alpha^{-1}\hat{x}_{i_{(A_\alpha)}})/\sum_{\alpha=1}^{F} d_\alpha^{-1}. \tag{5.28}$$

The largest number of analogues that are taken into account is F < N. Here F behaves like the "freedom-of-choice."

Alternatively, one can use a parametric inductive algorithm for combining the forecast analogues in which a complete polynomial of the form

$$\hat{x}_{i_{(B)}} = a_0 + a_1\hat{x}_{i_{(A_1)}} + a_2\hat{x}_{i_{(A_2)}} + \cdots + a_F\hat{x}_{i_{(A_F)}} \tag{5.29}$$

is used instead of the splines.

The following choices are to be considered to provide the most accurate forecasting process:

- choice of the optimal number of complexed analogues $F = F_{opt}$;
- choice of optimal set of features $m = m_{opt}$; and
- choice of the permissible variable measurement step width $h = h_{max}$.

Method of reducing variable set size

The two-stage method given below enables us to find the optimal set of effective features.

Stage 1. Variables are ordered according to their efficiency F = 1,2, 3, • • • (not more than five) using the partial cross- validation criterion $CVj \rightarrow \min$, defined with the help of moving a so-called "sliding window" (which is equal to one line) along the data sample (Figure 5.18). For each position of the "sliding line" its analogues are found in prehistory and the common analogue forecast is calculated using the splines. The discrepancy between the "sliding line" and the forecast analogue defines a forecast error for each variable. The error is found for all positions of the "sliding line" in the sample. The results are summed and averaged according to the following formulae:

$$\Delta\hat{x}_{ij} = |\hat{x}_{ij}(B) - \hat{x}_{ij}(A_1, A_2, \cdots, A_F)|$$

$$CV_j = \frac{1}{N}\sum_{i=1}^{N} |\Delta\hat{x}_{ij}|; \quad 1 \le j \le m.$$

$$CV_{\min} = \frac{1}{N}\sum_{i=1}^{N} |\Delta\hat{x}_{ij}|_{\min} \tag{5.30}$$

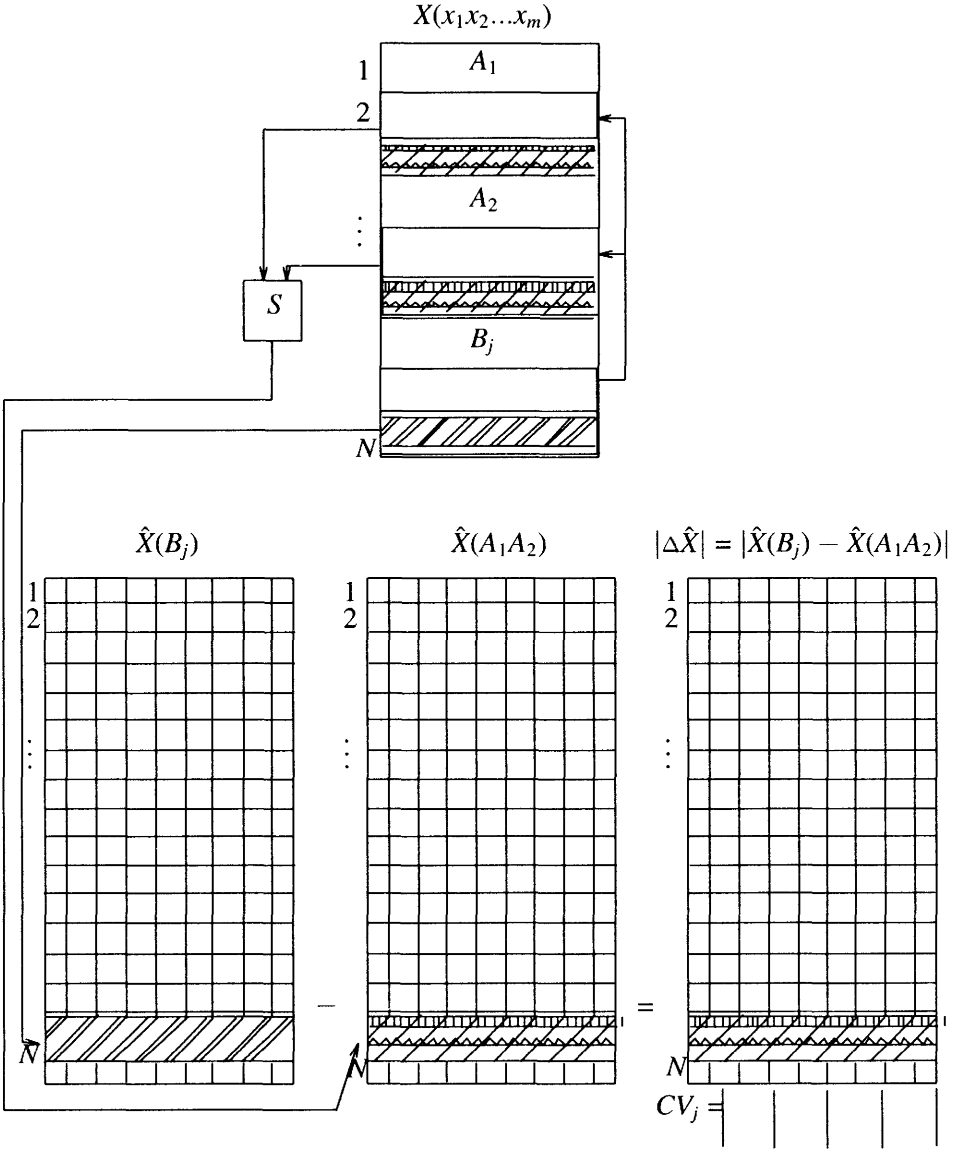


Figure 5.18. Schematic flow of the algorithm corresponding to process forecasts for calculating the cross-validation criterion CV_j when two analogues complexed, where fi-current position of sliding window, S-spline, and $|\Delta\hat{x}|$—absolute errors.

where *i,j* are numbers of data rows and columns respectively ($1 \leq i \leq N$, $1 \leq j \leq m$), $CV_{\min}$ is the cross-validation criterion for choosing optimal set of input variables (features), $|\Delta\hat{x}_{ij}|$ are the absolute values of errors, and $|\Delta\hat{x}_{ij}|_{\min}$ are the minimal value of $|\Delta\hat{x}_{ij}|$ in the lines of sample. In general, a different series of features ordered according to the criterion *CVj* are produced for different numbers *F* of complex analogues. This is analyzed on a plane of *F* versus *m*.

Stage 2. The feature series are arranged as per the values of the criterion *CVj.* A small number of feature sets are selected from all possible sets for further sorting out using the complete cross-validation criterion,

$$CV = \frac{1}{m}\sum_{j=1}^{m} CV_j \rightarrow \min. \tag{5.31}$$

The ordered feature set shows which sets should remain and which should be excluded. The complete set of feature sets is divided into groups, containing an equal number of features. Only one set, in which less efficient features are absent, remains in each group.

For example, there exists an ordered feature series of $x_3x_2x_1x_4$ (the best feature is x_3, the worst one is x_4); then the following sets are to be sorted out:

one set containing all four features: $x_3x_2x_1x_4$ (all included);
one set containing three features: $x_3x_2x_1$ (x_4 excluded);
one set containing two features: x_3x_2 (x_1x_4 excluded); and
one set consisting of one variable: x_3 ($x_2x_1x_4$ excluded).

The whole number of sets tested is equal to four, being equal to the number of features.

Algorithm for optimal forecast analogue

The schematic flow of mode of operation of the algorithm for optimal forecast analogue is illustrated in Figure 5.19. The overall algorithm consists of two levels: the first one corresponds to obtaining the optimal parameter set by using the two- stage method and the second one corresponds to the process forecasting. Figure 5.19a illustrates the analogue search $\hat{x}$ and evaluation of the forecast error δ for each position of the "sliding window" and the process observation. Figure 5.19b illustrates the efficiency estimation and ordering of variables using the criterion $CV_j \rightarrow \min$. Figure 5.19c illustrates how to obtain F_{opt} and m_{opt} with the help of the criterion $CV \rightarrow \min$. The variable sets are obtained using the criterion $CV_j \rightarrow \min$, and the complete cross-validation criterion $CV \rightarrow \min$ is calculated for them as explained above. The results are plotted on the plane of $F — m$, where the minimum value of the criterion is found. Optimization of the criterion for set of variables is evaluated as

$$\sum CV = \left(\frac{CV}{CV_{\min}} + \frac{CV_{\min}}{(CV_{\min})_{\min}}\right) \longrightarrow \max. \tag{5.32}$$

The point of the plane which gives the criterion minimum, defines the optimal parameters $F = F_{opt}$ and $m = m_{opt}$ sought for.

Variable set optimization enables the so called "useful" and "harmful" features in an initial sample to be highlighted; i.e., it makes possible the exclusion of some data sample columns. The forecast sought for is then read out from the sample using only those optimal parameter values. Figure 5.19d illustrates the forecast at the output position of the "sliding window" *B.*

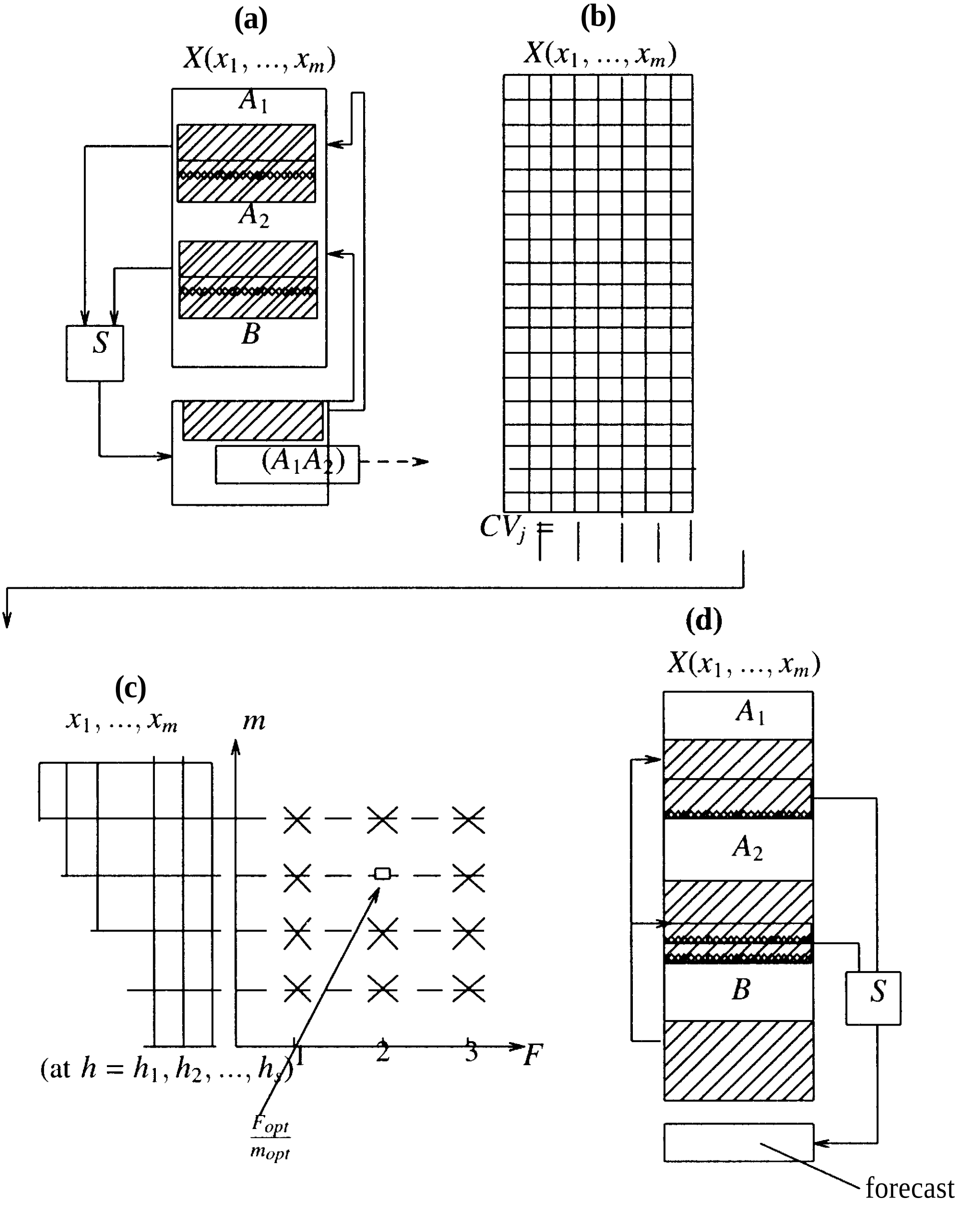


Figure 5.19. Modes of operation of recognition/forecast algorithm when two analogues A_1 and A_2 are complexed; (a) and (b) calculation of errors and criteria, (c) optimization of the criterion $\sum CV$, and (d) application mode; where $\sum CV = \left(\frac{CV}{CV_{\min}} + \frac{CV_{\min}}{(CV_{\min})_{\min}} \right) \longrightarrow \max$.

Pattern width optimization

This concerns the choice of permissible variable measurement width h_{max}. One observation point in the data table is called a pattern—in other words, it is a complete line of expansion. These lines of expansions can be transformed by summing up two, three, etc. adjoining lines and averaging the result. Due to overlapping of the number of lines in each junction, it is only reduced by unit; i.e., a sample containing twenty lines can be transformed into a sample containing nineteen doubled lines, or a sample containing eighteen tripled lines, and so forth. The sorting out of data sample makes it possible to select a permissible pattern width. Thus, the amount of sorting of the ensemble variants is reduced substantially if one succeeds in ranking the predictor-attributes (placing them in a row according to their effectiveness) in advance. The solution for the problem becomes simple. When the algorithm for optimal forecast analogue is used, one estimates each predictor separately according to the forecast measure $(\delta_i^2(x_i) \leq 1)$. This simplifies substantially the problem of choosing an effective ensemble of predictor-attributes. This means that one should identify the pattern width which provides a forecast variance value $\delta_i^2(x_i)$ less than unity for all variables treated. To estimate the value of $\delta_i^2(x_i)$, the forecast is to be calculated for the penultimate pattern.

We conclude that, in general, the optimization of the process forecast analogue algorithm is done in a three-dimensional space of the choices (F, m, h) for $Y = 0$, where F is the number of complexed analogues, m is the number of features taken into account, h is the data sample pattern width, and Y is the target function which is not specified.

5.2 Group analogues for event forecasting

The above procedure of process forecasting is described without specifying the output vector Y (target function); i.e., it deals only with the data sample of the variable attributes of X.

We extend this problem to a forecasting event where the output vector Y is defined as an event. In solving this type of problem, it is important that there be a correlation between the columns of the samples X and Y. However, it is usually absent. For successful events forecasting, samples X and Y must be complete and representative. In other words, the data sample has to contain a complete set of events of all types. For instance, when a crop harvest is forecasted, examples of "bad," "mean" and "good" harvests should be represented in Y. The data is complete if it contains a complete set of typical classes of observed functions. In addition, the sample should be representative. This means that clusters of matrices X and Y must coincide in time.

One of the tests for completeness and representativeness is that the matrices X and Y be subjected to cluster analysis using one of the known criteria. If identical correspondence clusters are obtained on the matrices (for example, good harvest has to correspond to good weather conditions and proper cultivation), then the sample is representative.

The problem of event forecasting is formulated in a more specific cause and effect manner and it has wider field of applications. In the formulation, the sample of attribute variables X is given in $(N+1)$time intervals, and the event factor Y is given in N intervals, if forecast of event Y in the $(N + 1)$st step is required. Some of the examples are:

1. **sample** X —observations of cultivation modes and weather conditions for $(N + 1)$ years.
   sample Y —harvest data for N years.
   It is necessary to predict the harvest for $(N + 1)$st year.
2. **sample** X —design and production features of $(N + 1)$ electronic devices.

sample Y—"life-time" and damage size data for N devices.
It is to predict the duration of uninterrupted operation of the $(N + 1)$st device.

3. to forecast the result of a surgical cancer treatment;
   sample Y—used as a loss vector containing three binary components; y_1 (recovery), y_2 (relapse), y_3 (metastases), and y_4 (the extent of disease, evaluated by the experts as a continuous quality).
   matrix X—includes various features (about 20), describing the state and method of surgical treatment for 31 patients.
   The results are known for 30 patients. These results are then used to predict the surgical treatment result for the recently operated 31st patient after the operation.

These are some typical examples of the event forecasting.

In order to predict the events, it is necessary to consider the following aspects to provide the accurate event forecasting;

- choice of the optimal number of complexed analogues $F = F_{opt}$;
- choice of the optimal set of features $m = m_{opt}$; and
- choice of the optimal target function vector $Y = Y_{opt}$.

The first two entities describe the process forecasting algorithm, whereas the latter is a specific aspect of the event forecasting problem.

Here, the pattern width (measurement step) h - 1 should not be changed. It is strictly equal to one line of an initial sample and the data sample cannot be transformed as explained before. Instead it is expedient to sort out the components of the vector Y (output value). For example, the harvest can be represented in the data sample not only by crops weight, but also by its sort and quality. The sorting out procedure allows only those components which give the minimal value for the criterion CV leading to a more accurate forecast to remain.

First, it is necessary to reduce the number of feature sets involved in the sorting. This is demonstrated in the Figure 5.20. The distinction from the method described in Figure 5.18 is that here two matrices X and Y are participating. Instead of getting the difference between sliding line and complexed analogue forecast, the differences of the vectors Y (not their forecasts) are calculated as

$$|\Delta x_{ij}| = x_{ij}(B) - x_{ij}(A_1, A_2, A_3)|. \tag{5.33}$$

The logic of feature choice is that the value of an effective feature at the current line and its analogues must be as close to each other as possible. A large discrepancy in the value means the feature does not define the output value Y; i.e., it is ineffective. The criterion CVj is calculated as the difference of feature values of the line, and the analogues averaged over the sample columns.

$$CV_j = \frac{1}{N}\sum_{i=1}^{N} |\Delta x_{ij}| \to \min. \tag{5.34}$$

Analogues are searched to find the matrix Y. At least one component of Y must be measured continuously and accurately for a unique analogue. However, if the analogue is not unique as defined, then the two components of a target function, which are derived from the Karhunen-Loeve algorithm, are added to the vector Y.

The schematic explanation to the algorithm is exhibited in Figure 5.21. Here "a" is the analogue choice, "b" is the calculation of the partial cross-validation criterion CVj → min for

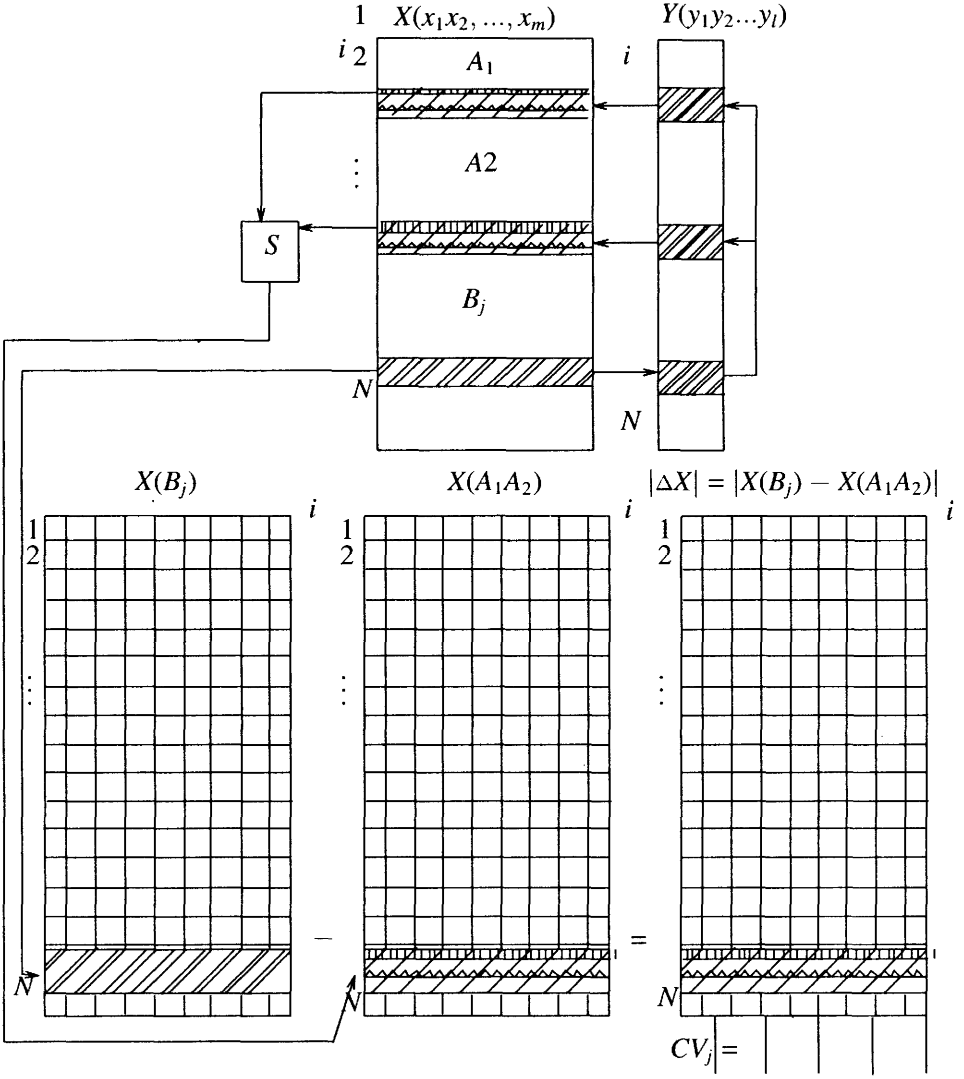


Figure 5.20. Schematic flow of the algorithm corresponding to events forecasts for calculating the cross-validation criterion CV_j when two analogues complexed as per the occurring events, where B-current position of sliding window, and S-spline; the criteria evaluated are $CV \rightarrow mm$, $CV_{\min} \rightarrow \min$, and $\sum CV = \left(\frac{CV}{CV_{\max}} + \frac{CV_{\min}}{(CV_{\min})_{\max}}\right) \longrightarrow \max$.

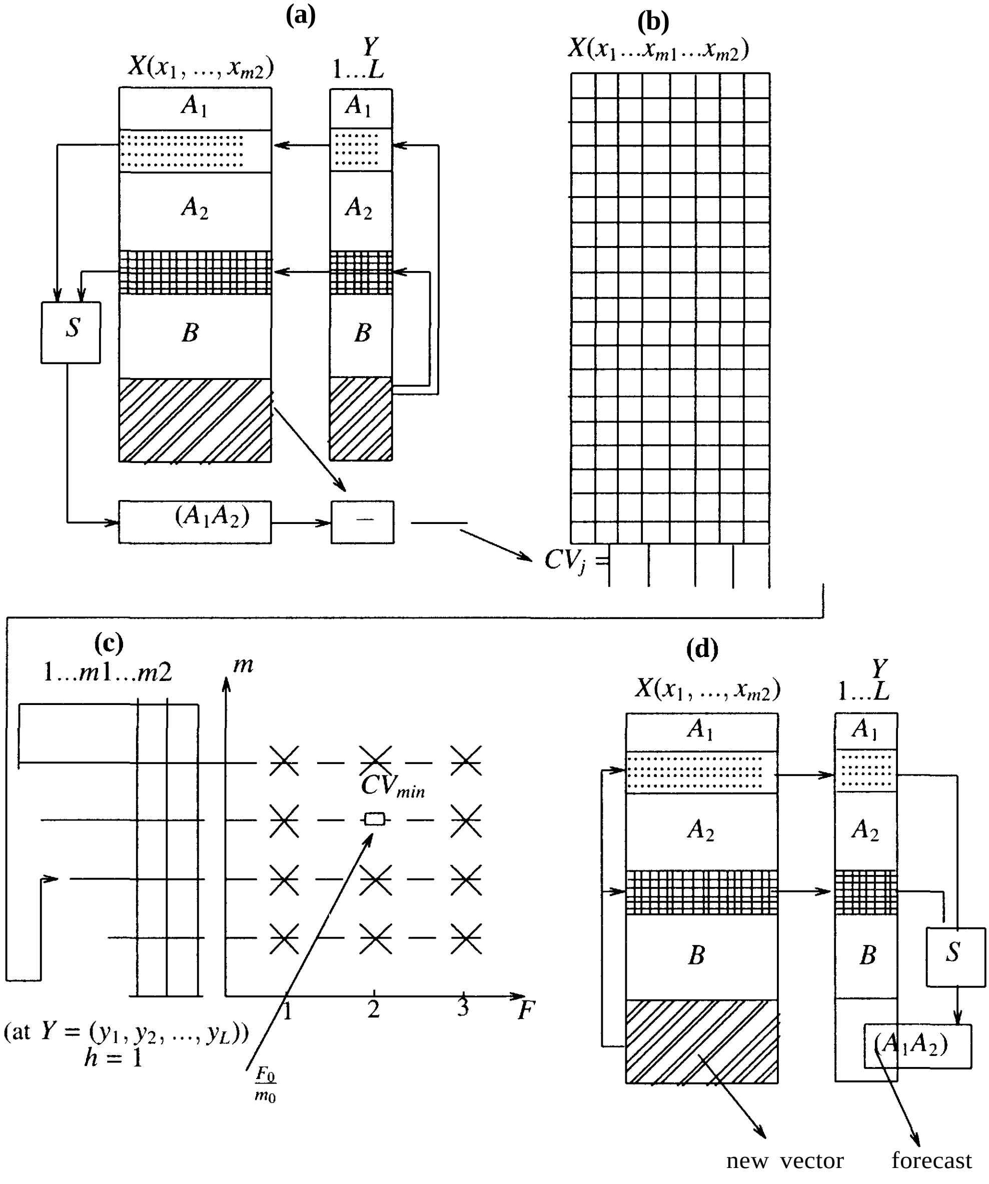


Figure 5.21. Modes of operation of an events forecast algorithm when two analogues $A\backslash$ and $A2$ complexed (a) choice of analogue, (b) calculation of partial cross-validation criterion CV_j, (c) arranging on the plane to obtain optimal point, and (d) the second stage of the event recognition/ forecast

ordering features, "c" is the calculation of values of the complete cross-validation criterion $\sum CV = \left(\frac{CV}{CV_{\max}} + \frac{CV_{\min}}{(CV_{\min})_{\max}}\right) \longrightarrow \max$. with the purpose of defining the optimal values of F_{opt} and m_{opt}. r is the forecast of the event corresponding to the $(N+$ l)st sample line under optimal algorithm parameter values.

Note that matrices X and Y are used in one direction (anti-clockwise) at the optimization stage, and in the opposite one (clockwise) at the forecast stage.

Otherfeatures

Use of convolution for an analogue choice in sorting out the vector components of Y. One can use a convolution of components in the target function instead of calculating the analogues in the multidimensional space. This helps the modeler to include components which lead to more accurate forecast. The analogues will be the same, but the calculations are simpler. The target function Y must have a continuous scale for a unique definition of the analogues. Thus, when at least one of the components of Y has such a reading scale, it is recommended that the convolution of the normalized component values $Y = \sqrt{(y_1^2 + y_2^2 + \cdots + y_l^2)}$ for analogue searching be used. If all components are binary variables (equal to 0 or 1), it is necessary to expand the component set by introducing one or two components of the orthogonal Karhunen-Loeve transform (for the joint sample XY).

$$Y = \sqrt{(y_1^2 + y_2^2 + \cdots + y_l^2 + z_1^2 + z_2^2)}, \tag{5.35}$$

where z\ and z_2 are components of the artificial target function [137]. Sorting out of the target function is meant for excluding some items from the expression.

The complete sorting of variants of criterion values $CV \longrightarrow \min$ is carried out in a three-dimensional space of (F, m, l) as $h = 1$, where F is the number of complex analogues, m is the number of feature sets, and / is the number of components in the target function.

Correlation measure of distances betweenpoints and "Wroslaw taxonomy." The simplest measure to calculate the distance between the points of the multidimensional feature space is the Euclidean distance for continuous features and Hamming distance for binary ones. If the data are nonstationary, for example, values will show an increasing or decreasing trend. The trend is then defined either as an averaged sum of normalized values of the variables or each variable trend is found separately (by a regression line in the form of polynomial of second- or third-order). Deviation of the variable from its trend is read out individually. The correlation coefficient of the deviation of each of the two measured points serves as a correlation measure of distance between them.

When the distance correlation measure is used, it is logical to apply the "Wroslaw taxonomy" algorithm for feature-ordering according to their efficiency. This algorithm is based on the partial cross-validation criterion $CV_j \rightarrow \min$ and makes it possible to order features according to their efficiency, and then excludes them one by one in the optimization process of the events-forecasting procedure to find the optimal feature set and the optimal number of complexed analogues.

The "Wroslaw taxonomy" algorithm is applicable only when the target function is defined in the problem. For this reason it is useful only in event forecasting, but not in the process forecasting.

Once the system is trained for a specific problem of event forecasting, it can be considered as the algorithm for recognition of new images. Thus, the event forecasting algorithm is treated as a particular case of the more general problem of image recognition; i.e., when recognizing the $(N+$ l)st vector of the target function Y is necessary.