Control and Cybernetics

Vol. 19 (1990) No. 1-2

Three-way clustering problems

by

JÓZEF POCIECHA
ANDRZEJ SOKOŁOWSKI

Institute of Statistics, Econometrics
and Computer Science
Academy of Economics
Rakowicka 27
31-510 Kraków, Poland

Three-way clustering problems (objects, attributes, time units) are discussed. Three types of clustering problems are defined. The notation and distinction of particular clustering problems is proposed. Some ways of transformation of the composite problems into simple problems is presented. Lidewise, some applications of proposed solutions in socio-economic investigations are shown in the paper.

1. Systematization of the three-way clustering problems

The reality to be analysed by clustering methods consists of a set of objects described by a set of variables, observed during some time span. Let us put:

$$Y = \{y_1, y_2, ..., y_m\} \text{ - the set of objects,}$$

$$Z = \{z_1, z_2, ..., z_w\} \text{ - the set of attributes,}$$

$$T = \{t_1, t_2, ..., t_n\} \text{ - the set of time units.}$$

For the simplicity of notation we are using the lower case letter without a subscript to denote a single element of the set.

The clustering problem is a relation which defines the way of creation of the set of operational taxonomic units (OTU's) and the classification space from Y, Z and T. We adopt two-position notation of clustering problems: $[\Omega, \Theta]$, where Ω is the set of OTU's and Θ is the classification space. Clustering approach implies that Ω should consist of at least two elements. The classification space can be created by sets Y, Z, T, their Cartesian products or Cartesian products of their

subsets. For further simplification we used the following notation for these products:
e.g.

$$ZY = Z \text{ x } Y = \{z_1y_1, z_2y_1, ..., z_wy_1, z_1y_2, z_2y_2, ..., z_wy_2, ..., z_wy_m\},$$

$$Yt = Y \text{ x } t = \{y_1t, y_2t, ..., y_mt\}.$$

The problem of classification of clustering problems has been discussed in Poland by Grabiński [1], Sokołowski [3] and Pociecha et al. [2].

1.1. Simple Problems

The particular sets Y, Z or T are the subjects of classification within this type of problems, so that these problems are defined according to the kind of OTU's.

1.1.1. Clustering of objects

$$[\,Y, zt\,] - \tag{1}$$

is the problem of clustering of single-attribute objects at a given time unit. The construction of structural series for a single variable can serve as the simplest example. In the field of statistical cluster analysis this problem can be described as the identification of elements in a mixture of one-dimensional distribution on the basis of a finite sample.

$$[\,Y, Zt\,] - \tag{2}$$

means classification of multidimensional objects in one time unit. This is the classical problem of cluster analysis.

$$[\,Y, zT\,] - \tag{3}$$

is the clustering of one-dimensional objects according to information from the time span T. This problem is important for the identification and estimation of a stochastic process. Suppose that we have the set of time series which are generated by one or more stochastic models. We are clustering these time series in order to define the number of generating models and to estimate expectations, standard deviations and autocorrelations for each stochastic process at each time point.

$$[\,Y, ZT\,] - \tag{4}$$

is the clustering of multidimensional objects within the time set T. This is the identification problem for mixture of multidimensional stochastic processes, see [2].

1.1.2. Periodizations

The set of time units is the subject of clustering within this type of problems. Periodization can lead to distinction of subsets of time units for which the objects were similar from the point of view of attributes (variables) used. The other result can be in the form of development phases each with relatively stable changes of objects.

$$[T, zy] \text{ -} \tag{5}$$

is the periodization of development of a single one-dimensional object, i.e. the partition of time series into segments. It is a problem of stationarity testing or searching for turning-points in time series. Usually some potential thresholds are shown by cluster analysis and then they are tested by means of comparing time series models or process parameters (mainly expected values).

$$[T, Zy] \text{ -} \tag{6}$$

is the periodization of the development of one object described by many variables.

$$[T, zY] \text{ -} \tag{7}$$

is the periodization of the set of one-dimensional objects. It is tantamount to stationarity testing for a one-dimensional stochastic process on a basis of set of observations, see [2].

$$[T, ZY] \text{ -} \tag{8}$$

is the periodization of the development of multidimensional object set.

Sometimes time continuity is required, i.e. "inside" units assigned to the particular phase there are no units from the other phases. Common clustering procedures usually do not obey this requirement but results can be easily adopted according to continuity principle.

1.1.3. Selection of diagnostic variables

The partition of the initial list of diagnostic variables into subsets is carried out. Selection of attributes plays an important role in cluster analysis for it has direct influence on final results of classification, particularly when equal weights for each variable are used.

$$[Z, yt] - \qquad (9)$$

is the classification of attributes for one object in one time unit. It is impossible to solve it because we have only one observation for each variable (differently named).

$$[Z, Yt] - \qquad (10)$$

is the dual problem for (2). Selection of independent variables for an econometric model also belongs to that group.

$$[Z, yT] - \qquad (11)$$

is the selection of attributes for the periodization of one object development. It is a dual problem to (6).

$$[Z, YT] - \qquad (12)$$

is the selection of attributes for the periodization of the set of objects.

1.2. Multiple Problems

The elements of sets generated by Cartesian products of Y, Z, T are the operational taxonomic units.

1.2.1. Classification in the attribute space

$$[YT, z] - \qquad (13)$$

is the periodization and classification of one-dimensional objects.

$$[YT, Z] - \qquad (14)$$

is the periodization and classification of multidimensional objects.

The above problems are the most interesting in the group of multiple ones. They are extensions of the classical problem through consideration of the time factor.

1.2.2. Classification in the object space

$$[ZT, y], \tag{15}$$

$$[ZT, Y]. \tag{16}$$

The solution of these problems gives, for instance, an answer as to which attributes and when take similar values in one-dimensional (15) or multidimensional (16) set of objects.

1.2.3. Classification in time space

$$[YZ, t], \tag{17}$$

$$[YZ, T]. \tag{18}$$

The Cartesian products of Y and Z are subjects of classification in time space, it means that single "object-variable" is understood as OTU.

1.3. Complex Problem

$$[YZT, .]. \tag{19}$$

This problem includes an arrangement of objects, variables and time units jointly.

The complete system of classification of three-way clustering problems, which has been presented in the paper, could be an easy tool for identification of clustering problems in empirical investigations.

2. Priority solutions of multiple and complex problems

The solution of multiple and complex problems gives many possibilities. Let us take the problem (13) or (14) of classification in attribute space. There are two ways of solving these problems:

a) without a priority — each one of "time-objects" is treated as single operational taxonomic unit and the clustering is being done for them as for ordinary "points". The resulting subsets are usually hard to explain because we do not a priori secure time or space continuity;

b) with a priority — we have to decide in advance whether we are interested in clustering of objects with an inclusion of time factor (space priority), or in periodization with a clustering of objects (time priority).

When setting assumptions on time priority we have to cluster the set of objects for each time unit. As the result we get n time-ordered partitions of the Y set:

$$\Delta_y : \{ \delta_t(Y) \} , \quad t \in T. \tag{20}$$

In other words Δ_y is the set of classifications of Y for each time unit.

Now there is the problem of how to calculate the distance matrix among classifications $\delta_t(Y)$. For each classification we define a binary membership matrix:

$$C = \{ c_{ij} \}; \quad i, j = 1, ..., m; \tag{21}$$

where:

$c_{ij} = 1$ — when objects y_i and y_j are in different subsets,
$c_{ij} = 0$ — when y_i and y_j are in the same subset.

Then for classifications l and k; $l, k \in T$, we calculate matrix D:

$$D = C_l + C_k , \tag{22}$$

From matrix D we calculate the dissimilarity measure of classifications:

$$p_{lk} = \frac{m^{(1)}}{m(m-1)} \quad , \tag{23}$$

where $m^{(1)}$ is the number of elements of matrix D equal to one. We calculate matrix P:

$$P = \{ p_{lk} \}; \quad l, k = 1, ..., n, \tag{24}$$

being the "distance matrix" among classifications $\delta_t(Y)$.

In the next step the set of time units is clustered in the classification space generated by Δ_Y. The problem can be defined as $[T, \Delta_Y]$. The development phases of the clustered set of objects are the final results.

The time priority in problem (14) can be described as:

$$[YT, Z] \Rightarrow [T, \Delta_Y], \delta_t(Y): [Y, Zt], \quad t \in T, \tag{25}$$

and the space priority as:

$$[YT, Z] \Rightarrow [Y, \Delta_T], \delta_Y(T): [T, Zy], \quad y \in Y, \tag{26}$$

where the "distance matrix" for $\delta_y(T)$ is calculated in similar way as above. The result of (26) gives the subsets of objects whose development was similar over T. With the time priority a typical ("average") partition of objects can be qualified for every development phase and with the space priority — a typical periodization for every subset of objects.

Within problem (16) time priority can be taken according to the following formula:

$$[ZT, Y] \Rightarrow [T, \Delta_Z], \delta_t(Z): [Z, tY], \quad t \in T, \tag{27}$$

or attribute priority according to the notation:

$$[ZT, Y] \Rightarrow [Z, \Delta_T], \delta_Z(T): [T, zY], \quad z \in Z \tag{28}$$

Attribute priority of classification in time space (18) can be defined as:

$$[YZ, T] \Rightarrow [Z, \Delta_Y], \delta_Z(Y): [Y, zT], \quad z \in Z \tag{29}$$

Applying the above formula it is possible to choose the attributes through direct estimatin of their delimitation ability for set Y in time T. Space priority follows:

$$[YZ, T] \Rightarrow [Y, \Delta_Z], \delta_y(Z): [Z, yT], \quad y \in Y \tag{30}$$

On the second stage of (30) the values of attributes are substituted by the classification of attributes performed separately for each object (in time T).

There are two principles in solving of complex problem (19). First without priority — where each element of the Cartesian product (time-objects-variables) is treated as a single OTU. On the other hand we can give the time priority as follows:

$$[YZT, .] \Rightarrow [T, \Delta_Y], \delta_t(Y): [Y, \Delta_Z t], \delta_t(Z): [Z, YT], \tag{31}$$

or object priority:

$$[YZT, .] \Rightarrow [Y, \Delta_T], \delta_y(T): [T, \Delta_Z y], \delta_y(Z): [Z, YT], \tag{32}$$

and try to solve these problems in the way presented above.

3. An example

The example comes from the field of international industrial comparisons. The set Y contains nine European countries:

1. Czecho-Slovakia	(CS)
2. France	(F)
3. East Germany	(DDR)
4. Poland	(PL)
5. West Germany	(D)

6. Rumunia	(R)
7. Hungary	(H)
8. Great Britain	(GB)
9. Italy	(I)

The set of attributes contains four variables of industrial production in natural units per capita:

1 — electricity production,
2 — steel production,
3 — cement production,
4 — sulfric acid production.

The time span contains seven years: 1980–1986.

First, the values of variables have been standardized, then Euclidean distances among countries in space of attributes have been calculated. Seven agglomerative clustering methods, see e.g. [2], have been applied:

— single linkage,
— complete linkage,
— average linkage,
— weighted average linkage,
— median,
— centroid,
— Ward's method.

As the best partition of OTU's, this partition was taken, for which the maximal jump of links in spanning tree is observed, see [2]. Some results of classifications for complete linkage method are shown below.

For year 1986 according to electricity production, the following clusters of countries have been obtained (clustering problem $[Y, z_1\ t_7]$):
four groups of countries —
(CS, F, DDR, PL, D), (R), (H, GB), (I).
For 1986, according to four main industrial attributes, results of clustering are, $[Y, Z\ t_7]$:
three groups of countries —
(CS, F, DDR, PL, D, R), (H), (GB, I).
Periodization of industrial development of Poland, $[T, Z\ y_4]$:
(1980, 81), (82, 83, 84), (85, 86).
Attributes' clustering for 1986, $[Z, Y\ t_7]$:
two groups —
(1, 2, 3), (4).

As an example of multiple problem the periodization and classification in multidimensional space has been presented, $[YT, Z]$. When we assume time priority we obtain periodization of the time period 1981-85.

I group — (1982), in which the typical partition of only socialist countries under consideration is:
(CS), (DDR, PL, R), (H).

II group — (1981, 83, 84, 85),
the typical partition of socialist countries in this group is:
(CS, DDR, PL, R), (H).

When we put object priority we obtain partition:

I group — (CS, DDR, PL), for which the typical periodization is:
(1981), (82, 83), (84, 85),

II group — (R, H), the typical periodization is:
(1981, 82, 83), (84), (85).

Within the complex problem according to formula (31) or (32) is such many concrete ways of solving this problem, that substantial economical interpretation of empirical calculations in this example becomes not so clear. For this reason we do not present results of this part of example.

References

[1] GRABIŃSKI T. Dynamiczne modele analizy taksonomicznej (Dynamic Models of Taxonomic Analysis, in Polish), AE, Kraków, 1975.

[2] POCIECHA J., PODOLEC B., SOKOŁOWSKI A., ZAJĄC K. Metody taksonomiczne w badaniach społeczno-ekonomicznych (Taxonomic Methods in Socio-Economic Research, in Polish), PWN, Warszawa, 1988.

[3] SOKOŁOWSKI A. O zagadnieniach taksonomicznych (On the Taxonomic Problems, in Polish), Zeszyty Naukowe AE w Krakowie, nr 165, 1982.

Received, March 1989.

Trójkierunkowe zagadnienia analizy skupień

W pracy rozważa się zagadnienia analizy skupień dla obiektów, atrybutów (zmiennych) i jednostek czasu. Wyróżnia się trzy typy takich zadań, dla których proponuje się zapis formalny i sposób wyodrębnienia. Przedstawiono sposoby przekształcania zadań złożonych w proste. Pokazano zastosowanie zaproponowanego formalizmu w badaniach społeczno-gospodarczych.

Три направления задач кластерного анализа

В работе расспатриваютця задачи кластерного анализа для объектов, атрибутов переменных и единиц времени. Выделяются три типа таких задач, для которых предлагается формальная запись и способ выделения. Представленны методы преобразования сложных задач в простые. Показаны применения предлагаемого формализма в общественно-экономических исследованиях.

Instructions for the Authors

"Control and Cybernetics" publishes original papers which have not been published and will not be simultaneously submitted essewhere. The preferred language of the papers is English.

Papers should not exceed 20 typewritten pages (210 × 297 mm) in length double spaced and with 500 mm margin on the left-hand side. Texts should be submitted in triplicate, typed only on one side of the sheet of paper.

The plan and form of the submitted typescripts is as follows:

1. The heading should include the title, full names and surnames of the authors in the alphabetic order, as well as the names and addresses of the institutions they represent. The heading should be followed by a concise summary (of approximately 15 typewritten lines).

2. Figures, photographs, tables, diagrams etc. should be enclosed with the typescript. The texts related should be typed on a separate page.

3. All elements of mathematical formulae should be typewritten whenever possible. A special attention is to be paid towards differentiating between capital and smaller letters. All the Greek letters appearing in the text should be defined. Indices and exponents should be written with special care. Round brackets should not be replaced by the inclined fraction line.

In general, elements easily confused are to be identified by the appropriate previously discussed measures or by a circled word or words explaining the element.

4. References should be listed in alphabetical order on a separate sheet. For journals the following information should appear: names (including initials or first names) of all authors, full title of paper, and journal name, volume, issue, pages, year of publication. Books cited should list author(s), full title, edition, place of publication, publisher, and year. Examples are:

Lukes D. Optimal regulation of nonlinear dynamical systems. SIAM J. Control 7 (1969), 1. 75-100.

Athans M., Falb P. Optimal Control. New York, Mc Graw-Hill 1966.

Wskazówki dla autorów

W wydawnictwie "Control and Cybernetics" drukuje się prace oryginalne nie publikowane w innych czasopismach. Zalecane jest nadsyłanie artykułów w języku angielskim. W przypadku nadesłania artykułu w języku polskim Redakcja może zalecić przetłumaczenie na język angielski. Objętość artykułu nie powinna przekraczać 1 arkusza wydawniczego, czyli ok. 20 stron maszynopisu formatu A4 z zachowaniem interlinii i marginesu 5 cm z lewej strony. Prace należy składać w 3 egzemplarzach. Układ pracy i forma powinny być dostosowane do niżej podanych wskazówek.

1. W nagłówku należy podać tytuł pracy, następnie imię (imiona) i nazwisko (nazwiska) autora (autorów) w porządku alfabetycznym oraz nazwę reprezentowanej instytucji i nazwę miasta. Po tytule należy umieścić krótkie streszczenie pracy (do 15 wierszy maszynopisu).

2. Materiał ilustracyjny powinien być dołączony na oddzielnych stronach. Podpisy pod rysunki należy podać oddzielnie.

3. Wzory i symbole powinny być wpisane na maszynie bardzo starannie.

Szczególną uwagę należy zwrócić na wyraźne zróżnicowanie małych i dużych liter. Litery greckie powinny być objaśnione na marginesie. Szczególnie dokładnie powinny być opisane indeksy (wskaźniki) i oznaczenia potęgowe. Należy stosować nawiasy okrągłe.

4. Spis literatury powinien być podany na końcu artykułu. Numery pozycji literatury w tekście zaopatruje się w nawiasy kwadratowe. Pozycje literatury powinny zawierać nazwisko autora (autorów) i pierwsze litery imion oraz dokładny tytuł pracy (w języku oryginału), a ponadto:

a) przy wydawnictwach zwartych (książki) — miejsce i rok wydania oraz wydawcę;

b) przy artykułach z czasopism: nazwę czasopisma, numer tomu, rok wydania i numer bieżący.

Pozycje literatury radzieckiej należy pisać alfabetem oryginalnym, czyli tzw. grażdanką.