A C T A U N I V E R S I T A T I S L O D Z I E N S I S
FOLIA OECONOMICA 286, 2013

*Justyna Wilk*[*], *Marcin Pełka*[**]

# CLUSTER ANALYSIS – SYMBOLIC VS. CLASSICAL DATA

**Abstract.** Clustering problem is addressed in many contexts and disciplines. Although there are numerous studies on cluster analysis, there is a lack of a review to complete and systematize knowledge of research approach depending on data form. The paper presents a concept of clustering, classifications of cluster analysis methods, comparison of numerical and symbolic taxonomy, specificity of symbolic data as regards classical data, methods of numerical and symbolic data analysis applicable in clustering procedure.

**Key words:** cluster analysis, symbolic data analysis, classification, numerical taxonomy, symbolic taxonomy.

## I. INTRODUCTION

Cluster analysis involves grouping of similar patterns to produce a classification. The clustering problem is addressed in many contexts and disciplines that reflects its broad appeal and usefulness as one of the most important methods of exploratory data analysis. One of the main application areas of cluster analysis is economic (regional, marketing, financial etc.) research such as e.g. market segmentation.

A complexity of economic problems and multiplicity of research approach require collecting data from various (primary and secondary) data sources for clustering purposes such as databases, questionnaire surveys etc. The data may take the form of classical or symbolic data. Additionally, clustering is a complex problem and its procedure consists of several stages which determine final results of an investigation. Selecting statistical methods and approach at each stage of the procedure is strictly determined by the processing data type.

Although there are numerous studies on cluster analysis, there is a lack of an overview study which would complete and systematize the knowledge of research approach depending on data form. The subject of this paper is to discuss statistical methods of numerical and symbolic data analysis which can be applied for clustering purposes.

[*] Ph.D., Chair of Econometrics and Computer Science, Wroclaw University of Economics.
[**] Ph.D., Chair of Econometrics and Computer Science, Wroclaw University of Economics.

The first part of the paper presents the aims of clustering, classifications of clustering methods with particular consideration of numerical and symbolic taxonomy. In the second section two types of data are distinguished and the concept and specificity of symbolic data is carefully discussed. The last part of this article identifies typical stages of clustering procedure and statistical methods designated for classical and symbolic data analysis.

## II. CLUSTER ANALYSIS

Cluster analysis consists in classification of patterns (objects, events) into relatively homogeneous groups based on a set of variables (characteristics, features). The main application areas of cluster analysis are: providing objects taxonomy, data reduction – objects (or variables) grouping of for further analyses, investigating similarities and dissimilarities between objects, confirmatory analysis of predefined hypothesis regarding data set structure (see Hair *et al.* (2006), pp. 561-562).

Methods applied for classification purposes discussed in this paper represent taxonomy approach, data mining, unsupervised learning, objects coincidence research, descriptive and non-parametric approach (Jain, Murty and Flynn (1999), Gatnar (1998), Anderberg (1973), Hair *et al.* (2006), Walesiak (2004), Koronacki and Ćwik (2005)). There is a multitude and diversity of clustering methods due to their properties (see Table 1).

One of the most important distinctions of cluster analysis algorithms, as regards the subject of this paper, is to distinguish methods of numerical and symbolic taxonomy (Table 2). This diversification results from research conducted in the area of symbolic data analysis (see e.g. Bock, Diday *et al.* (2000), Diday, Noirhomme-Fraiture *et al.* (2008)).

Table 1. Classification of cluster analysis methods

| Criterion | Groups of methods | Specificity |
|---|---|---|
| 1 | 2 | 3 |
| Starting point | agglomerative | Each object represents a separate class and the objects are joined together up to all of them belong to the same class |
| | divisive | All objects belong to the same class and a division procedure is taken up to each object represents a separate class |
| Classification results | hierachical | A clustering procedure results in a dendrogram (hierarchical tree) |
| | non-hierarchical | A clustering procedure results in a single division of set of objects |

Table 1 (cont.)

| 1 | 2 | 3 |
|---|---|---|
| Continuity of clustering process | iterative | An algorithm is based on successive iterations, e.g. moving objects between classes |
| | direct | A classification performs single operations route |
| Cluster membership | exclusive | The classes are separated; each object belongs to one class |
| | overlapping | The classes are not separated; each object may belong "fully" to one or more classes |
| | fuzzy | The classes are not separated; each object may belong partially to different classes |
| Data table | proximity matrix | Distance measurement between each pair of objects is required before clustering |
| | data table | Clustering procedure is based on data matrix or symbolic data table |
| Way of grouping | sequential | Classification requires repeated operation sequences |
| | simultaneous | Classification does not requires any repeated operation sequences |
| Grouping criterion | local | Optimization of division is performed separately at each stage of grouping |
| | global | There is one, the same optimization criterion at each stage of grouping |
| Data type | numerical taxonomy | Methods are designated to classical data form |
| | symbolic taxonomy | Methods are designated to symbolic data analysis |
| Number of classes | unknown | Number and quantity of clusters results from clustering process |
| | known | Number of clusters is specified before clustering but their quantity is known after completing the classification process |

Source: authors' elaboration based on Grabiński 1992; Gatnar 1998, Walesiak 2004, Wilk (2010a).

Table 2. Numerical vs. symbolic taxonomy

| Specification | Numerical taxonomy | Symbolic taxonomy |
|---|---|---|
| 1 | 2 | 3 |
| Theoretical background | – known classifications of: people (e.g. in India), plants and animals by Linneaeus (18th century), chemical elements by Mendelejew (19th century) etc.<br>– biology (biometrics) and anthropology research | – exploratory data analysis and data mining<br>– Artificial Intelligence and machine learning research<br>– applications of numerical taxonomy in biology<br>– cognitive psychology research<br>– conceptual clustering |
| Pioneering algorithm | – algorithm by Czekanowski (1913) | – EPAM algorithm by Fiegenbaum (1961) |
| Basic algorithms development | – the 50s and the 60s of the 20th century | – the 80s and the 90s of the 20th century |
| Designation | – classical data analysis | – symbolic data analysis |

Table 2 (cont.)

| 1 | 2 | 3 |
|---|---|---|
| Data table | – data matrix<br>– dissimilarity matrix | – symbolic data table<br>– dissimilarity matrix |
| Rules of classification | – grouping objects according to variables observations | – grouping similar objects to obtain useful class characteristics |
| Way of classification | – distances of objects and quality criterion are context-free measures<br>– distance of $A$ and $B$ object is the function of two objects: $d(A,B) = f(A',B')$ | – context-sensitive measures<br>– distance of $A$ and $B$ object can be presented as: $d(A,B) = f(A',B',O',C)$, where $O'$ – set of objects, $C$ – rules of classification |
| Quality criterion | – algorithm dependent | – mostly heuristic measures |
| Methods | – hierarchical, e.g. Ward's, complete linkage, centroid method<br>– nonhierarchical, e.g. $k$-means, $k$-medoids | – hierarchical, e.g. Brito's, Gowda and Diday's method<br>– nonhierarchical, e.g. SCULST, DCLUST, $k$-means by Verde |

Source: Authors' elaboration based on Gatnar (1998), Gordon (1999), Wilk (2010b).

## III. SYMBOLIC AND CLASSICAL DATA FORM

Numerical techniques were designed to investigate relations between objects understood as single individuals (e.g. persons, products, areas), described by quantitative (metric, numerical) and qualitative (non-metric, categorical) variables (see Table 3). An observation of each variable for the object results in a single value or category and a set of objects is presented in a data matrix.

Table 3. Types of classical variables

| Variable type | Measurement scale | Set of variable implementation | Main relations of variable implementation | Examples |
|---|---|---|---|---|
| 1 | 2 | 3 | 4 | 5 |
| Non-metric | Nominal | two and more disjoint (equivalent or mutually exclusive) categories | $y_A = y_B$, $y_A \neq y_B$ | sex, occupation, marital status, interest |
| | Ordinal | disjoint ordered categories or values (levels, ranks, grades, classes etc.) of relatively (non-valuated) comparisons | above and $y_A > y_B$, $y_A < y_B$ | level of education, social class, customer preference, product quality, level of satisfaction |

Table 3 (cont.)

| 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|
| Metric | Interval | single real numbers with contractual zero point and measure unit | above and $y_A - y_B = y_C - y_D$ | financial result, net migration, bank account balance, level of satisfaction on a scale [-100, 100] |
| | Ratio | single positive real numbers with natural (absolute) zero point and measure unit | above and $\frac{y_A}{y_B} = \frac{y_C}{y_D}$ | purchasing price, consumer income, age |

Source: Authors' elaboration based on Walesiak (1993), Mynarski (2000), pp. 79-83, Bock, Diday *et al.* (2000).

While numerical methods were dedicated to study relatively simple situations, the symbolic methods are designated to analyze symbolic data which are more complicated in their structure. Symbolic data analysis considers objects described by the variables whose implementations are in the form of intervals, set of categories, set of categories with weights and logically dependent structure (see Table 4). A special case of multivalued variable is non-metric variable as well as a special case of interval-valued variable is metric variable.

Table 4.Types of symbolic variables

| Variable type | Set of variable implementation | Main properties of variable implementation | Examples |
|---|---|---|---|
| Interval-valued | intervals of values | disjoint (ordered) intervals, non-disjoint intervals of real values | respondents' age, income; approximate price of product |
| Multivalued | set of values | set of categories (equivalent or ordered), real values, intervals of values | held driving license categories; components of products, knowledge of foreign languages |
| Modal | sets of values with weights | set of categories (equivalent or ordered), real values, intervals of values with associated weights (e.g. frequencies, probabilities) | proportional data of customers' expenses for food, clothes, services etc.; percentage share of population in regions by economic age groups |
| Dependent | hierarchic, logical, taxonomic structure of data | two and more (classical, symbolic) variables logically dependent | models and brands of cars, heights and weights of children, taxonomies of geographical regions |

Source: Authors' elaboration based on Bock, Diday *et al.* (2000), Diday, Noirhomme-Fraiture *et al.* (2008), Wilk (2010a).

Three types of symbolic objects as regard their specificity may be distinguished such as objects following classical approach, objects result from an aggregation of a set of objects described by classical variables and also synthetic objects result from describing properties of obtained clusters of first or second order symbolic objects (see Table 5). The set of observations referring to symbolic data is entered into a symbolic data table.

Table 5.Types of symbolic objects

| Object type | Specificity | Variables type | Examples |
|---|---|---|---|
| First order symbolic objects | objects following classical approach (primary units of the study) | classical variables, symbolic variables | respondent, territorial unit, product |
| Second order symbolic objects | result from the aggregation of a set of first order symbolic objects | symbolic variables | region made up of districts located in its territory |
| Synthetic objects | result from describing properties of obtained classes of first or second order symbolic objects | symbolic variables | market segment characteristics |

Source: Authors' elaboration based on Bock, Diday *et al.* (2000), Diday, Noirhomme-Fraiture *et al.* (2008).

The symbolic data form results from (see Bock, Diday *et al.* (2000), Diday, Noirhomme-Fraiture *et al.* (2008), Wilk (2010a), pp. 86-88):

1. Data nature, when an observation cannot be classified as a single value (imprecise, uncertain data), single category (conjunctive data) and independent data (e.g. taxonomies),
2. Surveys basing on questionnaire form with multiple choice questions (e.g. preferred brands of a product), sensitive personal information (e.g. customers' monthly expenditures), complex questions, e.g. place of residence (city: less than 100, 100-200, more than 200 thousand of inhabitants; village), linked questions (e.g. taxonomies, hierarchies),
3. The researcher's intention to aggregate collected data. Symbolic data results from classical data aggregation. The aggregation consists in the representation of lower order objects by means of higher order objects, e.g. lower level territorial units (e.g. NTS-4) into higher level territorial units (e.g. NTS-2), car versions (differ from acceleration, wheelbase, engine capacity, fuel type) into a model car, e.g. Volkswagen Golf. Such procedure is carried out to reduce a very large set of objects and also to refine the description of higher order objects, i.e. consider their internal structure (e.g. regional diversification of territorial unit).

## IV. CLUSTERING PROCEDURE

There are identified several steps that generally constitute a clustering procedure as follows (see Milligan (1996), pp. 342-343, Walesiak (2004), Gordon (1999), p. 7):

1. Objects and variables selection. Selecting research units and sometimes sampling is required. A set of relevant variables as regards a subject of the investigation that differentiated the set of objects must be chosen.
2. Variable normalization. If there are metric variables in the set it is usually necessary to unify their variability and dispose units of measurement.
3. Objects dissimilarity measurement. Proximity measurement is justified if selected cluster analysis method is based on distance matrix; the choice of the distance measure is strongly influenced by the nature of the data.
4. Objects classification. A selection of classification method is dependent on a subject of the investigation and quality of classification results.
5. Number of clusters' selection. It is provided by substantive knowledge of a researcher and sometimes also supported by formal algorithms.
6. Cluster validation. Assessing internal validity to reveal stability of cluster structure, its quality and robustness is usually supported by formal algorithms.
7. Cluster interpretation and profiling. Majority of empirical studies, apart from determining a number and quantity of clusters and objects membership, require defining cluster characteristics and distinguishing features.

This procedure is almost identical for clustering classical and symbolic data but the methods applied in each stage may differ (see Table 6). Classical data require applying methods developed in the area of numerical taxonomy, while symbolic data analysis is conducted using methods based on symbolic data table or dissimilarity matrix.

## V. CONCLUSIONS

Cluster analysis plays an important role in a wide variety of fields and is particularly useful in the area of economic research. It has evolved for decades to meet ongoing challenges. Developed solutions correspond to classical data situation, as well as symbolic data to analyze larger and larger data sets, and fuzzy, imprecise and conjunctive data.

Table 6. Clustering procedure – statistical methods of classical and symbolic data analysis

| Step | Classical data set | Symbolic data set |
|---|---|---|
| 1. Objects and variables selection | – methods designated for classical variables selection, e.g. HINoV method by Carmone, Kara and Maxwell | – methods of symbolic data analysis, e.g. Ichino's graph method, Talavera's method<br>– adjusted methods of classical data analysis, e.g. Carmone, Kara and Maxwell's HINoV method |
| 2. Variable normalization | – standardization, unitization, quotient transformation etc. | – |
| 3. Dissimilarity measurement | – Minkowski's metric, Mahalanobis distance, Walesiak's GDM, Sokal-Michener's measure | – distance measures of symbolic objects, e.g. Ichino-Yaguchi's, de Carvalho's, Gowda-Diday's |
| 4. Objects classification | – hierarchical methods of numerical taxonomy, e.g. Ward's method<br>– non-hierarchical methods of numerical taxonomy, e.g. *k*-means | – hierarchical methods of numerical and symbolic taxonomy, e.g. Ward's, Brito's method<br>– non-hierarchical methods of numerical and symbolic taxonomy, e.g. *k*-medoids, SCLUST |
| 5. Number of clusters' selection | – indices based on dissimilarity matrix, e.g. Baker i Hubert's, Hubert and Levine's<br>– indices based on data matrix, e.g. Caliński i Harabasz's | – indices based on dissimilarity matrix, e.g. Baker i Hubert's, Hubert and Levine's<br>– indices based on symbolic data table, e.g. Q(P) by Verde, Lechevallier and Chavent<br>– adjusted indices of numerical taxonomy, e.g. Caliński and Harabasz's |
| 6. Cluster validation | – methods of classical data analysis, e.g. Rousseeuw's silhouette index, replication analysis with Rand's index | – methods of symbolic data analysis, e.g. Bertrand–Bel-Mufti's method<br>– methods of classical data analysis, e.g. Rousseeuw's silhouette index<br>– adjusted methods of classical data analysis, e.g. replication analysis with Rand's index |
| 7. Cluster interpretation | – descriptive statistics | – Brito's CLINT technique |
| 8. Cluster profiling | – methods designated for classical data set such as classification trees (e.g. CART algorithm) and discrimination analysis | – methods of symbolic data analysis such as symbolic classification trees (e.g. TREE algorithm) and symbolic discrimination analysis |

Source: Authors' elaboration based on Bock, Diday *et al.* (2000), Diday, Noirhomme-Fraiture *et al.* (2008), Walesiak (2004), Everitt, Landau and Leese (2001), Gordon (1999), Wilk (2010a), Wilk and Pełka (2004).

## REFERENCES

Anderberg M.R. (1973), *Cluster Analysis for Applications*, Academic Press Inc., New York.
Bock H.H., Diday E. (Eds.) (2000), Analysis *of Symbolic Data. Exploratory Methods for Extracting Statistical Information from Complex Data*, Springer-Verlag, Berlin-Heidelberg.
Diday E., Noirhomme-Fraiture M. (Eds.) (2008), *Symbolic Data Analysis and the SODAS Software*, Wiley, Chichester.
Everitt B.S., Landau S., Leese M. (2001), *Cluster Analysis*, Fourth Edition, Arnold, London.
Gatnar E. (1998), *Symboliczne metody klasyfikacji danych*, PWN, Warszawa.
Gordon A.D. (1999), *Classification*, Chapman & Hall, London-New York-Washington.
Grabiński T. (1992), *Metody taksonometrii*, Wyd. AE w Krakowie, Kraków.
Hair J.F., Black W.C., Babin B.J, Anderson R.E., Tatham R.L. (2006), *Multivariate Data Analysis*, Pearson Prentice Hall, New Jersey.
Jain A.K., Murty M.N., Flynn P.J. (1999), Data Clustering: A Review, *ACM Computer Survey*, vol. 31, no. 3, pp. 264-323.
Koronacki J., Ćwik J. (2005), *Statystyczne systemy uczące się*, Wydawnictwa Naukowo-Techniczne, Warszawa.
Milligan G.W. (1996), Clustering Validation: Results and Implications for Applied Analyses, In: P. Arabie, L.J. Hubert, G. de Soete (Eds.), *Clustering and Classification*, World Scientific, Singapore, pp. 341-375.
Mynarski S. (2000), *Praktyczne metody analizy danych rynkowych i marketingowych*, Zakamycze, Kraków.
Walesiak M. (1993), Strategie postępowania w badaniach statystycznych w przypadku zbioru zmiennych mierzonych na skalach różnego typu, *Badania Operacyjne i Decyzje*, no. 1, pp. 71-77.
Walesiak M. (2004), Problemy decyzyjne w procesie klasyfikacji zbioru obiektów, In: J. Dziechciarz (Ed.), *Ekonometria 13. Zastosowania metod ilościowych*, PN AE we Wrocławiu, no. 1010, pp. 52-71.
Wilk J. (2010a), *Problemy segmentacji rynku z wykorzystaniem metod klasyfikacji i danych symbolicznych*, doctoral thesis, Jelenia Góra (unpublished).
Wilk J. (2010b), Cluster Analysis Methods in Symbolic Data Analysis, In: J. Pociecha (Ed.), *Data Analysis Methods in Economic Investigations*, Studia i Prace UE w Krakowie, no. 11, Kraków, pp. 39-54.
Wilk J., Pełka M. (2004), Dane symboliczne w zagadnieniu klasyfikacji, In: M. Rószkiewicz (Ed.), *Identyfikacja struktur rynkowych: pomiar – modelowanie – symulacja, Monografie i opracowania*, no. 533, Of. Wyd. SGH w Warszawie, Warszawa, pp. 103-120.

*Justyna Wilk, Marcin Pełka*

## ANALIZA SKUPIEŃ – DANE SYMBOLICZNE A DANE KLASYCZNE

Celem artykułu jest usystematyzowanie wiedzy na temat analizy skupień w zależności od rodzaju danych empirycznych opisujących problem badawczy. W artykule zaprezentowano cele analizy skupień, dokonano klasyfikacji metod analizy skupień, porównano metody taksonomii numerycznej i symbolicznej. Omówiono także specyfikę danych symbolicznych w odniesieniu do danych w ujęciu klasycznym oraz ich źródła w badaniach ekonomicznych. Wskazano metody statystyczne, jakie mają zastosowanie w analizie danych klasycznych i symbolicznych na każdym etapie procedury klasyfikacji.