# RECENT DEVELOPMENTS IN CLUSTER ANALYSIS

Anuška Ferligoj
Faculty of Social Sciences, University of Ljubljana
P.O. Box 47, 1109 Ljubljana, Slovenia
anuska.ferligoj@uni-lj.si

**Abstract:** *The purpose of cluster analysis is to investigate the structure within the set of units, in particular, to ask whether the units fall naturally into a certain smaller number of groups (or clusters, or classes) of units, such that units within a group are 'similar' to one another. Although such a clustering problem is intuitively simple and understandable, its solution is a continuing area of research. This is confirmed by the increasing number of papers on this and similar topics in the last three decades in journals of theoretical and applied statistics, the establishment of a special journal in 1984 for the field of cluster analysis, the Journal of Classification, and the foundation of the International Federation of Classification Societies in 1985. In the process of the development of cluster analysis, various types of clustering problems arose. In the paper the main topics are presented and some promising discussed.*

**Keywords:** algorithms, dissimilarities, consensus, sequence analysis, constraints, multicriteria clustering, conceptual clustering, symbolic clustering, cluster validity

## 1. Introduction

Grouping of units into clusters, so that the units within a cluster are as similar to each other as possible and the units between clusters dissimilar as much as possible is a very old problem and has been solved more or less skillfully for a number of years. Although such clustering problem is intuitively simple and understandable, its solving is very actual even today. This is confirmed by the increasing number of papers on this topic in the past decades in journals of theoretical and applied statistics, the establishment of a special journal in 1984 for the field of cluster analysis called *Journal of Classification*, and the foundation of the *International Federation of Classification Societies* in 1985. There are two main reasons for such an interest and developments in this field of data analysis:

- The clustering problems were, a few decades ago, solved separately in particular scientific fields, without having attempts to integrate specific solutions. This is a characteristic for the early stage in the development of a certain discipline. The first attempts of unifying different approaches for solving the clustering problems are seen in the sixties. The first more extensive work in this field, by Sokal and Sneath, appears in 1963. From then on, the field of cluster analysis develops as a specific discipline within data analysis.

- The development of cluster analysis was also strongly influenced by the development of computer technology. The computers allowed the application of more demanding computational procedures and the processing of large data sets. Theoretical results in computer science are also important, especially the results of the computational complexity theory. It has been proven just fifteen years ago that most of the clustering problems are NP-hard. Therefore it is not surprising that they were and still are being solved with heuristic approaches, more or less adapted to the specifics of the particular problem.

In the process of the development of cluster analysis, various types of problems arose. In this paper some promising subareas are presented and discussed.

## 2. General clustering problem

Cluster analysis (also classification, taxonomy) deals mainly with the following general problem: given a set of units $E$, determine subsets of it, called clusters $C$, which are homogeneous and/or well separated. Clustering $\mathcal{C}$ is a set of clusters. It can be a partition or a more complicated structure as a hierarchy. Clustering problem can be formulated as an optimization problem:

> Determine the clustering $\mathcal{C}^*$ for which
>
> $$P(\mathcal{C}^*) = \min_{\mathcal{C} \in \Phi} P(\mathcal{C})$$
>
> where $\mathcal{C}$ is a clustering of a given *set of units or actors* $E$, $\Phi$ is the set of all possible clusterings and $P : \Phi \to \mathbb{R}$ a *criterion function.*

Such criterion functions can be constructed *indirectly* as a function of a suitable (dis)similarity measure between pairs of units (e.g., Euclidean distance) or *directly.* In most cases the criterion function is defined indirectly. In the case of partitions into k clusters the Ward criterion function is usually used

$$P(\mathcal{C}) = \sum_{C \in \mathcal{C}} \sum_{X \in C} d(X, T_C)$$

where $T_C$ is the centroid of the cluster $C$ and $d$ the squared Euclidean distance.

The usual steps of solving a clustering problem are (Hansen, Jaumard, Sanlaville 1993):

1. Select the set of units $E$.
2. Observe or measure appropriate variables according to the given problem. Variables can be measured on different types of scale.
3. Choose an appropriate dissimilarity according to the given problem and type of measured variables.
4. Choose an appropriate type of clustering (e.g., partition, hierarchy, pyramid).
5. Select or create an appropriate criterion function to evaluate the selected type of clustering (e.g., Ward criterion function).
6. Choose or devise an algorithm for the given clustering problem.
7. Determine the clustering(s) which optimize(s) the chosen criterion with the selected algorithm. Only an approximate solution can sometimes be found due to the lack of en exact algorithm, or to excessive computing time for reaching an exact solution.
8. Apply various tests to detect whether the obtained solutions has some underlying structure or not. Use descriptive statistics to summarize the characteristics of each cluster.

In general, the clustering problems are NP-hard problems. This is the reason why different efficient heuristic algorithms producing 'good' clustering solutions are found.

Most of the statistical systems as SAS, SPSS and BMDP have implemented the hierarchical and leader algorithm.

The scheme of the hierarchical algorithm is the following one:

Each unit is a cluster: $C_i = \{X_i\}$ , $X_i \in E$ , $i = 1, 2, \ldots, n$;
**repeat** while there exist at least two clusters:
determine the nearest pair of clusters $C_p$ and $C_q$:
$d(C_p, C_q) = \min_{u,v} d(C_u, C_v)$ ;
fuse clusters $C_p$ and $C_q$ into a new cluster $C_r = C_p \cup C_q$;
replace the clusters $C_p$ and $C_q$ by the cluster $C_r$;
determine the dissimilarities between the cluster $C_r$
and other clusters according to the selected method
(e.g., single linkage, complete linkage, Ward).

The resulting clustering (hierarchy) can be graphically represented by the clustering tree or dendrogram.

The scheme of the leader algorithm, or K-MEANS or dynamic clusters method is the following one:

determine the initial set of leaders $\mathcal{L} = \{L_i\}$
**repeat**
determine the clustering $\mathcal{C}$ in such a way to classify
each unit to the nearest leader,
for each cluster $C_i \in \mathcal{C}$ compute its center $\bar{C}_i$.
The center $\bar{C}_i$ determines the new leader $L_i$ of the cluster $C_i$;
**till** the leaders do not change.

There are many other clustering algorithms, as e.g., relocation algorithm, graph algorithms.

## 3. Recent developments in cluster analysis

Arabie and Hubert (1992) in their recent review considered the following new subareas of intense development of cluster analysis: clustering of binary data, measures of association or dissimilarity coefficients, mixture models, overlapping clustering, partitioning, constrained clustering, consensus clustering, cluster validity, variable selection and weighting, computational advances, and substantive developments.

At the Conference of the International Classification Societies (IFCS'93), which was held in Paris, Hans Bock (1993) also reviewed various achievements in cluster analysis and pointed to a series of unresolved problems. He discussed the following subareas: formal characterization of cluster concepts, probabilistic clustering models and optimum partitions, validity of classifications and significance testing, distance-based and graph-theoretical methods, clustering of variables, model selection methods, computational and software aspects, new data types: conceptual and symbolic clustering, and retrieval-oriented classification methods.

Another interesting insight on new developments in cluster analysis gives an overview of titles of sessions presented at the IFCS'93 and IFCS'96 conferences: dissimilarities,

distances and classification, sequence analysis, classification and trees, consensus, analysis of vague data, uncertainty and classification, probabilistic clustering models, symbolic cluster analysis, comparing classifications, constrained clustering, spatial clustering, classification in biology, and phylogeny. There was another conference on the analysis of dissimilarities DISTANCIA '92 in 1992 in which very similar topics were also discussed.

In the following subsections some of the most promising are discussed in more detail.

### 3.1 (Dis)similarities

There are several issues of similarity measures treated in data analysis. When choosing a similarity measure various elements have to be considered: its mathematical properties, its behavior when confronted with data sets, the nature of the data, the use that will be made of the dissimilarity matrix etc. Gower and Legendre (1986) discussed metric, Euclidean and some other properties of dissimilarities and also how the information obtained may be used to guide the choice of a dissimilarity in particular applications.

The theoretical foundations of similarity measures lay on measurement theory. Several authors (e.g., Critchley and Van Cutsem 1992) are working on the order theoretic foundations of clustering theory. Critchley and Van Cutsem obtained a very general result establishing bijections between types of dissimilarities and corresponding clustering structures and their representations. Equivalences and dissimilarities between similarity measures are also studied (Batagelj and Bren 1995).

Dissimilarities can be also represented in different ways (e.g., by multidimensional scaling, graphs). An early overview of these representations was given by Blumental (1953). The importance of this subarea of cluster analysis was also shown when the International Federation of Classification Societies decided to give the first prize for young researchers to Klauer for his work on representation of dissimilarities by graphs (e.g., Klauer and Carroll 1991). Bandelt and Dress (1992) proposed additive decomposition theory for distance matrices, which can be applied to phylogenetic analysis. The 'split decomposition' of a given distance matrix results in a number of weighted splits that can be represented by graphs.

It is also important to mention that appropriate dissimilarity measures should be developed for specific problems and type of data, as for example in the case of sequences and symbolic objects. These two special cases are discussed in more detail below.

### 3.2 Comparison and consensus of clustering

There is often the case, that several clusterings should be compared. Many indices have been proposed to compare partitions. Among them Hubert and Arabie (1985) approached this problem indirectly by assessing the congruence of two proximity matrices using a simple cross-product measure. Leclerc (1985), Benkaraache and Van Cutsem (1993) and others proposed indices to compare hierarchical clusterings.

On the other hand many formal developments relevant to the consensus of clusterings can be found in the last thirty years. Among first were Regnier (1965) and Adams (1972) developed consensus methods for partitions and hierarchical clusterings. Boorman and Olivier (1973) employed metrics on lattices to compare clusterings. After these first steps an enormous development in this field arose. Many consensus methods for clusterings were proposed (e.g., McMorris and Neuman 1983) and indices of consensus among clusterings (e.g., Day 1983). Barthelemy and Monjardet (1981, 1988) developed a unified treatment of median procedures for consensus problems. A special issue of Journal of Classification (Day 1986) was devoted to comparison and consensus of clusterings. Excellent reviews

of this topic are also available (e.g., Faith 1988; Leclerc 1988). Numerous axiomatic frameworks have been devised for consensus structures (e.g., Barthelemy and Janowitz 1991).

### 3.3 Sequence analysis

William Day stated in his plenary lecture at the IFCS'93 Conference, that "in biology the problem of comparing nucleic or amino acid sequences is of considerable practical importance. When properly aligned, multiple sequences may be used to estimate evolutionary relationships among the organisms they present, or to obtain a summary of them as a consensus sequence. Even when sequences exhibit little overall similarity, the presence of short shared patterns or motifs may indicate distant familial relationships or biologically significant molecular features. Nor is sequence comparison a problem restricted to molecular biology, for it finds application in speech research, the editing of text files, and even the analysis of bird songs. Perhaps it is not surprising that a rich and varied literature on sequence comparison has developed since the early seventies, one containing hundreds of theoretical or methodological contributions."In recent years, researchers have begun to use consensus techniques for analyzing molecular sequences. Day and McMorris (1992) critically compared such consensus methods. Several papers discussed and proposed new approaches to identifying consensus in molecular sequences also at the IFCS'93 and IFCS'96 conferences.

### 3.4 Constrained clustering

In the case of constrained clustering, grouping of similar units into clusters according to the selected characteristics has to satisfy also some additional conditions. These problems are also relatively old. One of the most frequently treated problem in this field is regionalization: clusters of similar geographical regions have to be found according to the chosen characteristics, by which the regions, which compose the cluster, have to be geographically connected. In the literature, there have appeared a number of approaches to the analytical determination of regionalization. Majority of authors (e.g., Lebart 1978; Ferligoj and Batagelj 1983; Gordon 1987; Legendre 1987) solved this problem by adapting standard clustering procedures, especially the agglomerative hierarchical algorithm and the local optimization procedures (e.g., the relocation algorithm), that while determining the clustering they also test whether the units which compose the clusters satisfy the additional condition of geographical contiguity. This type of constrained clustering is also called clustering with relational constraint. Ferligoj and Batagelj (1982, 1983) first treated the clustering problem with relational constraints for the general symmetric and later also for the nonsymmetric relation (an example of a nonsymmetric relation is friendship). A review of clustering with the symmetric relational constraint approaches was given by Murtagh (1984).

It is also possible to find some other, non-relational conditions - clustering constraints in the literature. Such an example is the constraint on a given (constraining) variable. Clustering with a constraining variable has been dealt with by several authors (e.g., Ferligoj 1986).

The third group of constraints given special attention were the optimizational demands while clustering. Such clustering problems were solved above all with the threshold methods (Lefkovich 1985). Another, probably more promising approach is multicriteria clustering approach (Ferligoj and Batagelj 1992) where one criterion is defined by a selected

clustering criterion and others by optimizational constraints.

### 3.5 Multicriteria clustering

Some clustering problems cannot be appropriately solved with classical clustering algorithms because they require optimization over more than one criterion. In general, solutions optimal according to each particular criterion are not identical. Thus, the problem arises of how to find the best solution so as to satisfy as many as possible of all the criteria considered. In this sense the set of *Pareto efficient* clusterings is defined: a clustering is Pareto efficient if it cannot be improved on any criterion without sacrificing some other criterion.

A multicriteria clustering problem can be approached in different ways:

- by reduction to a clustering problem with a single criterion obtained as a combination of the given criteria; or
- by consensus clustering techniques applied to clusterings obtained by single criterion clustering algorithms for each criterion; or
- by constrained clustering algorithms where a selected criterion is considered as the clustering criterion and all others determine the constraints; or
- by direct algorithms. Hanani (1979) proposed an algorithm based on the dynamic clusters method using the concept of the kernel, as a representation of any given criterion. Ferligoj and Batagelj (1992) propose modified relocation algorithms and modified agglomerative hierarchical algorithms.

### 3.6 Conceptual and symbolic clustering

In contrast of objects described by classical data matrices (the rows represented individuals and the columns variables) or (dis)similarity matrices, modern information systems are concerned with objects like verbal texts, documents, chemical structures and images, which are primarily described in conceptual, symbolic, linguistic or semantic terms (Bock 1993). Cluster analysis provides interesting techniques for analysis and handling such objects. This research leads to symbolic data analysis (e.g., Diday 1984), conceptual, logical or knowledge-based clustering (e.g., Diday 1990; Wille 1992) etc. Without too much risk we can conclude, that conceptual and symbolic clustering is one of the most promising area in the future development in cluster analysis.

### 3.7 Cluster validity

As different clustering methods can generally produce different solutions on the same data, the question arises whether the clusters have "reality"or validity vis-a-vis the data (Arabie and Hubert 1992). Jain and Dubes (1988) distinguish three strategies for validation: *external criteria*, measuring performance by matching a clustering structure to a priori information, *internal criteria* , assessing the fit between the structure and the data, using only the data themselves, and *relative criteria*, deciding which of two structures is better in some sense, such as being more stable or appropriate for the data. Arabie and Hubert (1992) mention that among the issues most commonly investigated are selection of indices of cluster structure and their distributions and determining the appropriate number of clusters. The second issue was discussed by several authors also at the last two

IFCS conferences and we can conclude, that this is still among the difficult problems in cluster analysis.

### 4. Conclusion

In the paper an attempt to present the most promising developments in cluster analysis is given. Cluster analysis is still understood as a descriptive technique whose results are too dependent upon the effects of used methods. Therefore several research problems which are not yet adequately solved should be studied in the near future. Among them inferential aspects in cluster analysis and the insights of existing clustering algorithms. Special attention should also be given to linking the gap between clustering theory and practice of using clustering methods.

## References


[1] Adams, E.N., III (1972): Consensus techniques and the comparison of taxonomic trees. *Systematic Zoology*, 21, 390-397.

[2] Arabie P., and Hubert L.J. (1992): Combinatorial data analysis. *Annu. rev. Psychol.*, 43, 169-203.

[3] Bandelt, H.J., and Dress, A. (1992); A canonical decomposition theory for matrices on a finite set. *Advances in Mathematics*.

[4] Barthelemy, J.P., and Monjardet, B. (1981): The median procedure in cluster analysis and social choice theory. *Mathematical Social Sciences*, 1, 235-267.

[5] Barthelemy, J.P., and Monjardet, B. (1988): The median procedure in data analysis: New results and open problems. In: Bock, H.H. (Ed.): *Classification and Related Methods of Data Analysis*, Amsterdam: North-Holland, 309-316.

[6] Barthelemy, L.P., and Janowitz, M.F. (1991): A formal theory of consensus. *SIAM J. Disc. Math.*, 4, 305-322.

[7] Batagelj, V., and Bren, M. (1995), Comparing resemplance measures. *Journal of Classification*, 12, 73-90.

[8] Benkaraache, T., and Van Cutsem, B. (1993): Comparisons of hierarchical classifications. Paper presented at the IFCS-93, Paris, August 31 - September 4, 1993.

[9] Boorman, S.A., and Olivier, D.C. (1973): Metrics on spaces of finite trees. *Journal of mathematical Psychology*, 10, 26-59.

[10] Bock H.H. (1993): Classification and clustering: Problems for the future. In: *IFCS-93 Abstracts*, Paris: TELECOM, 133-134.

[11] Blumenthal, L.M. (1953): *Theory and Applications of Distance Geometry*. London: Oxford University Press.

[12] Critchley, F., and Van Cutsem, B. (1992): An order-theoretic unification and generalisation of certain fundamental bijections in mathematical classification. I and II. *Research Reports*, No. 874 and 875, Grenoble: IMAG and LMC.

[13] Day, W.H.E. (1983): The role of complexity in comparing classifications. *Mathematical Biosciences*, 66, 97-114.

[14] Day, W.H.E. (Ed.) (1986), Consensus classification [Special issue], *Journal of Classification*, 3, (2).

[15] Day, W.H.E., and McMorris, F.R. (1992): Critical comparison of consensus methods for molecular sequences. *Nucleic Acid Research*, 20, 1093-1099.

[16] Diday, E. (1984): Une representation visuelle des classes empietantes: Les piramydes. RAIRO, *Analyse des donnees*, 52, 475-526.

[17] Diday, E. (1990): Knowledge representation and symbolic data analysis. In: M. Schader and W. Gaul (Eds.): *Knowledge Data and Computer Assisted Decisions*, Berlin: Springer-Verlag, 17-34.

[18] Faith, D.P. (1988): Consensus applications in the biological sciences. In: Bock, H.H. (Ed.): *Classification and Related Methods of Data Analysis*, Amsterdam: North-Holland, 325-332.

[19] Ferligoj A. (1986): Clustering with constraining variable. *Journal of Mathematical Sociology*, 12, 299-313.

[20] Ferligoj A., and Batagelj V. (1982): Clustering with relational constraint. *Psychometrika*, 47, 413-426.

[21] Ferligoj A., and Batagelj V. (1983): Some types of clustering with relational constraint. *Psychometrika*, 48, 541-552.

[22] Ferligoj A., and Batagelj V. (1992): Direct multicriteria clustering algorithms. *Journal of Classification*, 9, 43-61.

[23] Gordon A.D. (1973): Classification in the presence of constraints. *Biometrics*, 2, 821-827.

[24] Gordon A.D. (1980): Methods of constrained classification. In: R. Tomassone (Ed.): *Analyse de Donnee Informatique*. INRIA, Le Chesnay.

[25] Gordon A.D. (1987): Classification and assignment in soil science. *Soil use and management*, 3, 3-8.

[26] Gower, J.C., and Legendre, P. (1986): Metric and Euclidean properties of dissimilarity coefficients. *Journal of Classification*, 3, 5-48.

[27] Hanani, U. (1979): *Multicriteria Dynamic Clustering*, Rapport de Recherche No. 358, Rocquencourt: IRIA.

[28] Hansen P., Jaumard B., and Sanlaville E. (1993); Partitioning problems in cluster analysis: A review of mathematical programming approaches. Invited paper presented at the IFCS-93, Paris, August 31 - September 4, 1993.

[29] Hubert, L.J., and Arabie, P. (1985): Comparing partitions. *Journal of Classification*, 2, 193-218.

[30] Jain, A.K., and Dubes, R.C. (1988): *Algorithms for Clustering Data*. New York: Prentice-Hall.

[31] Klauer, K.C., and Carroll, J.D. (1991): A comparison of two approaches to fitting direct graphs to nonsymmetric proximity measures. *Journal of Classification*, 8, 251-268.

[32] Lebart L. (1978): Programme d' Agregation avec Contraintes (CAH Contiguit). *Les Cahiers d'Analyse des Donnes*, 3, 275-287

[33] Leclerc, B. (1985): La comparaison des hierarchies: Indices et metriques. *Math. Sc. Hum.*, 92, 5-40.

[34] Leclerc, B. (1988): Consensus applications in the social sciences. In: Bock, H.H. (Ed.): *Classification and Related Methods of Data Analysis*, Amsterdam: North-Holland, 333-340.

[35] Lefkovitch L.P. (1985): Multi-criteria clustering in genotype - environment interaction problems. *Theoretical and Applied Genetics*, 70, 585-589.

[36] Legendre P. (1987): Constrained clustering. In: P. Legendre and L. Legendre (Eds.): *Developments in Numerical Ecology*, Berlin: Springer-Verlag, 289-307.

[37] McMorris, F.R., and Neuman, D. (1983): Consensus functions defined on trees. *Mathematical Social sciences*, 4, 131-136.

[38] Murtagh F. (1984): A survey of algorithms for contiguity- constrained clustering and related problems. *The Computer Journal.*

[39] Regnier, S. (1965): Sur quelques aspects mathematiques des problems de classification automatique. *I.C.C. Bulletin*, 4, 175-191.

[40] Sokal R.R., and Sneath P.H.A. (1963): *Principles of Numerical Taxonomy*. Freeman, San Francisco.

[41] Wille, R. (1992): Concept lattices and conceptual knowledge systems. *Computers and Mathematics with Applications*, 23, 493-515.