

# Quality Criteria for GMDH-based Clustering

Lyudmyla Sarycheva

Dept. of Geoinformation Systems, National Mining University, 49005, K. Marx .av., 19, Dnipropetrovsk, Ukraine

sarycheval@nmu.org.ua

**Abstract.** *New external criteria of clustering quality estimation (GMDH-criteria) are proposed. The GMDH-criteria are based on splitting of initial sample  $\mathbf{X}$  containing  $n$  clustered objects, in two not intersected equivalent subsamples A and B. Each subsample A for an object corresponds to a subsample B of the object. They form together the pair named a dipole. The GMDH-criteria generate minimum in area of underfitted clusterizations and allow to find clusterization of optimal complexity in a case of noisy data.*

## Keywords

Group Method of Data Handling (GMDH), external criteria of clustering quality estimation, modeling, clusterization

## 1 Introduction

A variety of cluster analysis algorithms results to situation when the same data generate, commonly, various classifications. Therefore it is required to obtain validation of the structure, which the clustering brings to the data. It is necessary to analyze such properties of clusters as density, dispersion, size, form, separability.

Unequivocal quantitative characteristics of these properties are absent in the literature. The majority of universal programs of the data analysis (Statistica, Statistical Toolbox MatLab, SPSS, Data Mining) offer a broad spectrum of the cluster analysis methods but have not the procedures of quality check for the obtained solution.

A look at clustering as modelling allows to transfer basic concepts of the GMDH theory of models self-organizing to the cluster analysis theory and methods.

Self-organizing of clusterizations means their exhaustive search for optimum selection. The more discrepancy of the data, the easier is the optimum clusterization (the complexity is determined by number of clusters and number of features). In algorithms of objective cluster analysis, clusters are created by an internal criterion and their optimum number and structure of an ensemble of features are determined by an external criterion (minimum in area of underfitted clustering, optimal for a given level of disturbance dispersion).

Generally accepted approach to clustering based on the accuracy criteria is effective only for precise and complete input data. If the disturbances are absent, all criteria (both external and internal) indicate true clusterization [1]. For the noisy data, it is necessary to find the clusterization of optimal complexity (i.e. noncontradictory clusterization) using the external criteria.

**The purpose of this article** is to develop external criteria of clusterization quality based on GMDH principles.

## 2 Internal Clusterization Quality Criteria

Let  $x_{ij}$  be measurements of features describing given set of objects - observations  $X$  ( $i = 1, 2, \dots, n$  is number of observations,  $n$  is a quantity of observations,  $j = 1, 2, \dots, m$  is a number of feature,  $m$  is a feature index). Input data represent an "object - feature" matrix  $(x_{ij})$ :  $\mathbf{X}^j = (x_{1j}, x_{2j}, \dots, x_{nj})^T$  is a column vector of  $j$ -th feature values for  $n$  objects,  $\mathbf{X}_i = (x_{i1}, x_{i2}, \dots, x_{im})$  is a row vector of  $m$  indices values of  $i$ -th object.

The clusterization  $K = \{K_1, K_2, \dots, K_k\}$ ,  $1 \leq k \leq n$  of ensemble  $X$  is the set of nonempty, mutually nonintersecting subsets (clusters)  $K_q$ ,  $q = 1, 2, \dots, k$ , of the ensemble  $X$ , which combination coincides with  $X$ :

$$K_1 \cup K_2 \cup \dots \cup K_k = X; K_i \cap K_j = \emptyset, i \neq j; K_q \neq \emptyset, i, j, q = 1, 2, \dots, k.$$

The clusterization  $K^* \subseteq \Phi$  is the best, if

$$K^* = \arg \max_{K \subseteq \Phi} J(K) \quad (\text{or } K^* = \arg \min_{K \subseteq \Phi} J(K)),$$

where  $\Phi$  is an ensemble of all permissible splittings (clusterizations) of given ensemble  $X$ ;  $J(K)$  is a criterion of clusterization quality. The number of clusters  $k$  can be unknown beforehand.

Any method of clusterization has the internal criterion. The majority of known methods of clusterization is based on using internal accuracy criteria or information criteria. The following internal criteria of clusterization quality estimation are most widespread:

1) criterion of intracluster dispersions (used in the  $k$ -means method)

$$J_1 = \sum_{j=1}^k \sum_{\mathbf{X}_i \in K_j} d_E^2(\mathbf{X}_i, \boldsymbol{\mu}_j),$$

where  $\boldsymbol{\mu}_j = \frac{1}{n_j} \sum_{\mathbf{X}_i \in K_j} \mathbf{X}_i$  is a barycenter of cluster  $K_j$ ;  $n_j$  is a number of objects in it;

2) criterion of paired intracluster distances between objects

$$J_2 = \sum_{j=1}^k \frac{1}{n_j} \sum_{\mathbf{X}_i, \mathbf{X}_g \in K_j} d_E^2(\mathbf{X}_i, \mathbf{X}_g);$$

3) criterion of an intercluster scatter of objects (the more magnitude of  $J_3$  ( $0 < J_3 < 1$ ), the greater portion of a common objects scatter is illustrated by the intergroup scatter and the quality of a partition is better)

$$J_3 = 1 - \frac{W}{S},$$

where  $W = \sum_{j=1}^k W_j$ ;  $W_j = \sum_{\mathbf{X}_i \in K_j} d^2(\mathbf{X}_i, \boldsymbol{\mu}_j)$  is an intracluster scatter;  $S = \sum_{i=1}^n d^2(\mathbf{X}_i, \bar{\mathbf{X}})$  is a common dispersion,

$\bar{\mathbf{X}} = \frac{1}{n} \sum_{i=1}^n \mathbf{X}_i$  – common barycentre.

4) generalized intracluster dispersion:

$$J_4 = \det \left( \sum_{j=1}^k n_j \mathbf{C}_j \right) \quad \text{or} \quad J_4^* = \prod_{j=1}^k (\det \mathbf{C}_j)^{n_j},$$

where  $\det(\mathbf{C})$  is determinant of a matrix  $\mathbf{C}$ ;  $\mathbf{C}_j$  is a covariance matrix of a cluster  $K_j$ .

### 3 Quality Criteria for GMDH-based Clustering

In articles [2-6] the definitions of internal and external GMDH-criteria are given, the classification of external criteria is given.

The commonality of principles of models and clusters self-organizing allows defining external criteria of clusterization (similar to criteria of accuracy, consistency and others) for research of clusters validation depending on the data used.

To substantiate a way of partitioning and selecting criterion of clusterization quality estimation we shall consider a hypothetical situation, when  $k^o$  and structure of objectively existing clusters  $K_1, K_2, \dots, K_{k^o}$  are known, i.e. the objective clusterization is known.

Let's enter the suppositions:

$$\text{H1. } r_{\max} < R_{\min},$$

$$\text{H2. } n_g = 2l \quad \forall g = 1, 2, \dots, k^o,$$

where  $r_{\max} = \max_q r_q$ ,  $q \in \{1, 2, \dots, k^o\}$ ;  $n_g$  is a number of objects in a cluster  $K_g$ ;

$r_q = \max_{\mathbf{X}_i, \mathbf{X}_k \in K_q} d(\mathbf{X}_i, \mathbf{X}_k)$  is a distance between the remote objects in a cluster  $K_q$ ;

$R_{\min} = \min_{g, q} R_{g, q}$ ,  $g \neq q$ ;  $g, q \in \{1, 2, \dots, k^o\}$ ;

$R_{g, q} = d(K_g, K_q)$  is a distance between clusters  $K_g$  and  $K_q$ .

We will split initial sample  $X$  containing  $n$  clustered objects into two not intersected equivalent subsamples  $A$  and  $B$  with dimensions  $\frac{n}{2} \times m$ ,  $A \cap B = \emptyset$ ,  $A \cup B = X$ ,  $X^T = [X_A^T; X_B^T]$ .

A) Compute  $n(n-1)/2$  distances  $d(\mathbf{X}_i, \mathbf{X}_j)$  between objects  $\mathbf{X}_i$  and  $\mathbf{X}_j$ ,  $i = 1, 2, \dots, n-1$ ;  $j = i+1, i+2, \dots, n$ .

B) Define objects  $\mathbf{X}_q$  and  $\mathbf{X}_s$  such, that  $d(\mathbf{X}_q, \mathbf{X}_s) = \min_{i, j} d(\mathbf{X}_i, \mathbf{X}_j)$ .

C) The  $\mathbf{X}_q$  object is entered in subsample  $A$ , and  $\mathbf{X}_s$  object, nearest to it, in subsample  $B$ . Each subsample  $A$  of the object corresponds to a subsample  $B$  of the object. They together form the pair  $(\mathbf{X}_q, \mathbf{X}_s)$  named a dipole.

D) Repeat steps B) - C) for the residual objects and distances between them, while all objects will be entered in  $A$  and  $B$ . The subsample  $A$  contains objects with numbers  $q_1, q_2, \dots, q_{n/2}$ , and the subsample  $B$  contains objects with numbers  $s_1, s_2, \dots, s_{n/2}$  ( $n$  it is supposed even, in case of odd  $n$  any one object of last pair is considered twice).

Let's conduct at the same time clusterization for subsamples  $B$  and  $A$  and calculate the sum of intracluster distances between objects on subsample  $B$  by results of clusterization on  $A$  and on the contrary same magnitude on  $A$  by results of clusterization on  $B$  (correspondence of objects subsamples  $A$  and  $B$  is established by their membership to one dipole):

$$J_{AB} = \sum_{q=1}^{k_A} \sum_{i_q^A, j_q^A}^{n_q^A} d^B(i_q^A, j_q^A) + \sum_{q=1}^{k_B} \sum_{i_q^B, j_q^B}^{n_q^B} d^A(i_q^B, j_q^B), \quad (1)$$

where  $q$  is a number of a cluster;

$k_A$  is current number of clusters in subsample  $A$ ;  $k_B$  – is current number of clusters in subsample  $B$ ;

$i_q^A, j_q^A$  are numbers of objects in subsample  $A$ ;  $i_q^B, j_q^B$  – are numbers of objects in subsample  $B$ ;

$n_q^A$  is current number of objects in a cluster  $K_q$  in subsample  $A$ ;

$n_q^B$  is current number of objects in a cluster  $Q_q$  in subsample  $B$ ;

$d^B(i_q^A, j_q^A)$  is a distance between two objects in subsample  $B$ , first of which has in pair object  $i_q^A$ , and second has object  $j_q^A$ ;

$d^A(i_q^B, j_q^B)$  is a distance between two objects in subsample  $A$ , first of which has in pair object  $i_q^B$ , and second has object  $j_q^B$ .

It is possible to prove that if as measure of likeness between two objects an Euclidean distance is taken and as measure of a likeness between object and cluster or between two clusters a nearest-neighbor distance is taken, clusterization will be carried out on agglomerative hierarchical algorithm, and recalculation formula for distances from joined cluster  $S_1 \cup S_2$  up to other clusters  $S$  is fixed as

$$d(S, S_1 \cup S_2) = \frac{1}{2} (d(S, S_1) + d(S, S_2) - |d(S, S_1) - d(S, S_2)|),$$

the suppositions H1 and H2 are carried out, the criterion  $J_{AB}$  (1) has a minimum at  $k = k^0$ .

Similarly to exterior criteria GMDH for models [6], using the sum of paired intracluster distances between objects as  $RSS$ , it is possible to define exterior criterion for clusterization:

$$J_{RS} = \frac{J_X - J_{AB}}{J_{AB}}, \quad (2)$$

where  $J_X$  is a sum of paired intracluster distances between objects on sampling  $X = A \cup B$ .

To find clusterization, in which the centers conforming to each other on subsamples  $A$  and  $B$  clusters are matched, the criterion may be used:

$$J_R = \frac{1}{k \cdot m} \sum_{i=1}^k \sum_{j=1}^m (\bar{x}_{ij}^A - \bar{x}_{ij}^B)^2, \quad (3)$$

where  $k = k_A = k_B$  is current number of clusters in subsamples  $A$  and  $B$ ;  $m$  is a number of coordinates,

$\bar{x}_{ij}^A, \bar{x}_{ij}^B$  are  $j$ -coordinates of centers of  $i$ -th clusters constructed on  $A$  and  $B$ .

Let  $K = \{K_1, K_2, \dots, K_{k_A}\}$  is clusterization for subsample  $A$ , and  $Q = \{Q_1, Q_2, \dots, Q_{k_B}\}$  is clusterization for subsample  $B$ . The criterion of a consistency of clusterizations constructed on subsamples  $A$  and  $B$ , looks like:

$$J_H = \frac{\frac{1}{2} (\sum_{i=1}^{k_A} |K_i|^2 + \sum_{i=1}^{k_B} |Q_i|^2) - \sum_{i=1}^{k_A} \sum_{j=1}^{k_B} |K_i \cap Q_j|^2}{\frac{1}{2} (\sum_{i=1}^{k_A} |K_i|^2 + \sum_{i=1}^{k_B} |Q_i|^2)}, \quad (4)$$

where  $k_A, k_B$  is a number of clusters (subsets of the initial ensemble) in clusterizations  $K$  and  $Q$  accordingly;

$|K_i|, |Q_j|, i = 1, 2, \dots, k_A; j = 1, 2, \dots, k_B$  are potencies of appropriate subsets, i.e. elements number in clusters  $K_i$  and  $Q_j$ .

Value  $J_H$  accepts values from 0 up to 1:  $J_H = 0$  in case of completely coincident partitions in clusterizations  $K$  and  $Q$ ;  $J_H = 1$  when at completely distinct.

Objective clusterization is  $K^o = \{K_1, K_2, \dots, K_{k^o}\}, 1 < k^o < n$ , for which the conditions are fulfilled:

$$\begin{aligned} & 1) r_{\max} < R_{\min}, \\ & 2) K^o = \arg \min_{K \subseteq \Phi} J_{AB}(K), \end{aligned} \quad (5)$$

where  $J_{AB}(K)$  is the weighted sum of external criteria (1), (4).

For objective clusterization the greatest number of objects of pairs  $(q_l, s_l), l = 1, 2, \dots, n/2$ , are contained in the conforming clusters for subsamples  $A$  and  $B$ . For example, if the objects with numbers  $q_3, q_7, q_{10}$  have hitted in one cluster for subsample  $A$ , and the objects with numbers  $s_3, s_7, s_{10}$  have hitted in one cluster for subsample  $B$ ; and clusters structure on  $A$  and  $B$  coincides (number of clusters, number of objects in the conforming clusters on  $A$  and  $B$  identical, and the pairs  $(q_l, s_l), l = 1, 2, \dots, n/2$ , are in the conforming clusters), this clusterization is objective (fig.1).

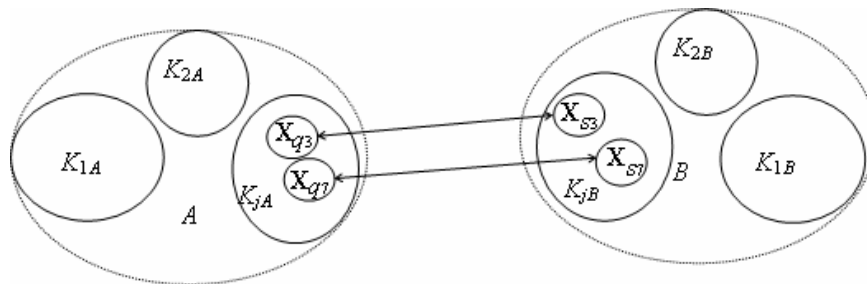


Fig. 1. Correspondence of clusterizations for subsamples  $A$  and  $B$

It is impossible to check conditions H1 and H2 in real environment, therefore to find an objective clusterization one should use interactive scheme, when for each clusterization-candidate the check of conditions H1 and H2 by the obtained outcomes is fulfilled. Taking into account that «the exhaustive search of the clusterizations-candidates in essence does not differ from exhaustive search of a set of the models-candidates» [2], the pluriserial iterative algorithm GMDH is used for finding objective clusterization [7].

## 4 Experiments

Advantages of the proposed clusterization quality criteria are demonstrated on the two data sets.

**I. Fisher's Irises.** The iris data published by Fisher (1936) are widely used for examples in cluster analysis. The sepal length, sepal width, petal length, and petal width are measured in millimeters on 150 iris specimens from each of **three** species, *Iris setosa*, *Iris versicolor*, and *Iris virginica* ( $n=150, m=4, k=3$ ).

It's appropriate to represent clusterization results in diagram form (fig. 2), namely depict clusterization quality criterions' values  $J$  on plots, where the value range is measured at  $Oz$ -direction; the number  $k$  of clusters is measured at  $Ox$ -direction (from 1 to 6), and the number  $m$  of features at  $Oy$ -direction (from 2 to 4). The GMDH-criteria generate minimum in area  $k=3$ .

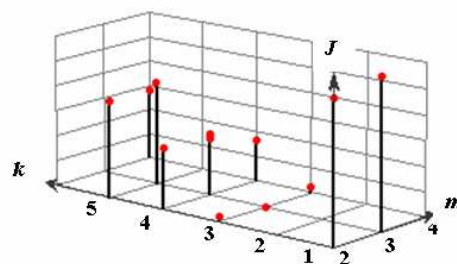
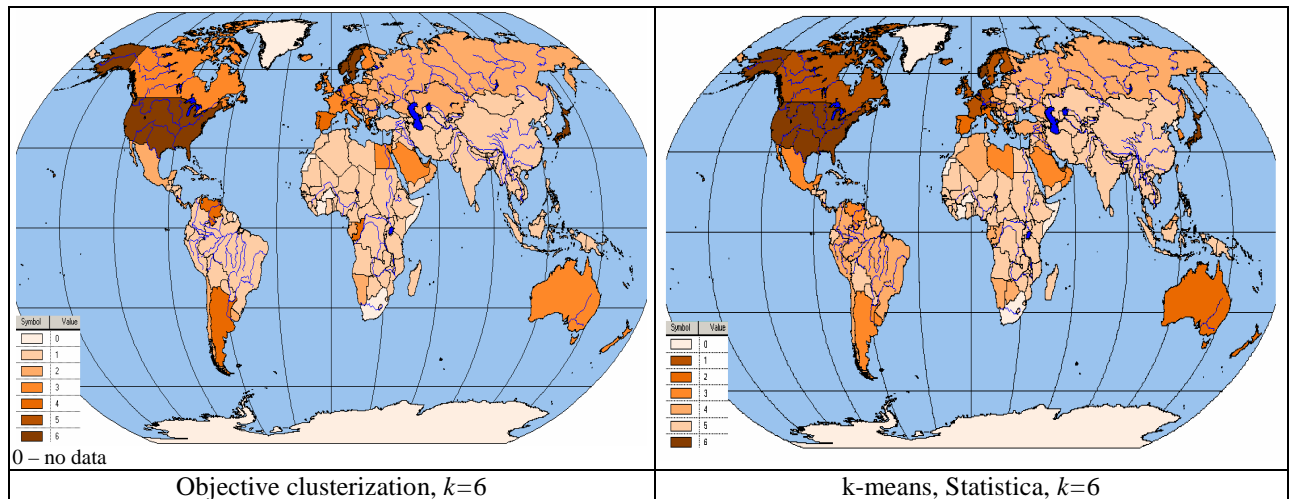


Fig. 2. Diagram of clusterization quality GMDH-criterions

**II. Parameters of gross domestic product (GDP)** measured during 40 years (1967-2007) for 192 countries of the world (ERS International Macroeconomic Data Set).

Let's represent the values of matrix, which shows region belonging to this or that class, as a geoiconic model is range maps for visual comparison analysis of experimentally obtained results (fig. 3). Clusterization has been carried out for numbers of clusters from 2 to 10, but as example the clusterization for  $k=6$  is shown (the GMDH-criteria generate minimum in area  $k=6$ ).



**Fig. 3.** Clusterization results

## 5 Conclusions

New external criteria of clustering quality estimation (GMDH-criteria) are proposed. The GMDH-criteria are based on splitting of initial sample  $X$  containing  $n$  clustered objects, in two not intersecting equivalent subsamples  $A$  and  $B$ . Each subsample  $A$  for an object corresponds to a subsample  $B$  of the object. They form together the pair named a dipole. The clustering both for subsamples  $B$  and  $A$  is carried out simultaneously and the sum of intercluster distances for sample  $B$  by results of clustering of sample  $A$  and the same value for  $A$  by results of clustering of  $B$  is calculated.

The GMDH-criteria generate minimum in area of underfitted clusterizations and allow to find clusterization of optimal complexity in a case of noisy data.

## References

- [1] A.G. Ivakhnenko. Objective clusterization based on the theory of self-organizing of models // Automation, 1987.– №5.–P.6-15. (In Russian)
- [2] A.G. Ivakhnenko. Method of series test (exhaustive search) of the clusterizations-candidates by criteria of a differential type // Pattern recognition, classification, prognosis. Mathematical methods and their application. B. 2. – Moscow: Nauka, 1989. – P.126-158. (In Russian)
- [3] A.G. Ivakhnenko and V.S. Stepashko, Pomekhoustoychivost' modelirovaniya (Noise Immunity of Modelling), Naukova dumka, Kiev, 1985.– 216 p.
- [4] A.G. Ivakhnenko. Induktivnyi metod samoorganizatsii modelei slozhnykh sistem (An Induktive Method of Self-Organization of Models of Complex Systems). – Kiev: Naukova Dumka, 1987. – 296 p. (In Russian)
- [5] A.G. Ivakhnenko. Continuity and discretization. Exhaustive search methods of simulation and clusterization.. – Naukova dumka, Kiev, 1990. – 224 p. (In Russian)
- [6] V.S. Stepashko, Yu. L. Kocherga. Methods and criterion of problem solving of structural identification // Automation, 1985. – № 5. – P. 29–37. (In Russian)
- [7] A.G. Ivakhnenko and Yu. P. Yurachkovskiy, Modelirovaniye slozhnykh system po eksperimental'nym dannym (Modelling of Complex Systems for Experimental Data), Radio I Svyaz', Moscow, 1987.– 120 p. (In Russian)