The GMDH Cluster Analysis Model*

HE Chang-zheng[1], XU Xiao-zhan[2], XIAO Jin[1]

[1]Business School, Sichuan University; [2.] College of Mathematics, Sichuan University, #24 South, First Section, NO.1 Ring Road, Chengdu 610064, P.R. China

changzhenghe911@hotmail.com xuxzmail@163.com xjxiaojin@126.com

Abstract: *The method of the objective cluster analysis is appraised by analyzing its algorithm steps. A new consistency criterion is proposed and the method of the GMDH cluster analysis is established. We show, through theoretical analysis and demonstration comparisons, that the GMDH cluster analysis is the further development of the objective cluster analysis.*

Keywords: *cluster analysis, consistency criterion, objective cluster analysis, GMDH cluster analysis.*

1 Introduction

Cluster analysis has found wide applications in many fields such as pattern recognition, data analysis, image processing and so on. In market research, cluster analysis can help market analysts to find different customer groups from the customer database and characterize the feature of different customer groups through their purchase pattern. Cluster analysis has already become an active research topic in the area of data mining.

There are different methods of cluster analysis. However, they all have different requirements for a priori knowledge of the system[1]. For example, the partitioning method requires the partition numbers to be given beforehand, which amounts to asking the user to input needed parameters. Since the clustering results are sensitive to the input parameters, this will increase the burden of the user and makes it difficult to control the clustering quality. As another example, the hierarchical method requires the model-maker to select a distance level from the dendrogram obtained from a priori knowledge of the system. This requirement makes the clustering results rely excessively on a priori knowledge of the model-maker.

Dr A.G. Ivakhnenko, academician of the Ukrainian Academy of Sciences, applied the core concept and principles of GMDH (the Group Method of Data Handling)[2] to clustering, thus creating a new cluster analysis method—the Objective Cluster Analysis (known as OCA)[3]. This new method can automatically and objectively determine the number of clusters and find the optimal clustering scheme[4].

Based on the analysis of the consistency criterion of OCA, this paper proposes a new consistency criterion. The new clustering method, using the new consistency criterion, will be called the GMDH Cluster Analysis(or GCA for short).

* This work is supported by Natural Science Foundation of China (70271073).

2 Basic steps of OCA

In order to look for the optimal number of the clusters, OCA first uses "dipoles" to divide sample data into two subsets, *A* and *B*. Then it evaluates the consistency of clustering schemes on sets A and B with consistency criterion η_c [5]. Let *m* be the number of variables of the objects to be clustered, *n* be the number of the whole samples. Then we have the measurable sample

$$X^T = (x_1, x_2, ..., x_n), \text{ where } x_i = (x_{i1}, x_{i2}, ..., x_{im}), (i = 1, 2, ..., n).$$

Now the basic steps of OCA are as follows[6]:

2.1 Compute the distance between samples x_i and x_j

$$d_{ij} = \sqrt{\sum_{k=1}^{m} (x_{ik} - x_{jk})^2}, \quad i = 1, 2, ..., n, \quad j = i+1, i+2, ..., n.$$

2.2 Partition the data samples

C_n^2 dipoles in the form $\begin{pmatrix} x_i \\ x_j \end{pmatrix}$ can be obtained from *n* data samples, and let d_{ij} in Section 2.1 will be called the value of the dipole $\begin{pmatrix} x_i \\ x_j \end{pmatrix}$. These dipoles are arranged in ascending order of their values. Then we take the first $k = [n/2]$ (where $[\cdot]$ is the integral function) dipoles with no common samples. Let *A* be the subset of all x_i's and *B* be the subset of all x_j's of these dipoles. Thus by using "dipoles" we divide evenly the set of data samples into two subsets *A* and *B* (see Section 2.3).

After deleting those dipoles that generated the subsets *A* and *B*, sample subsets *C* and *D*, which are used as testing set, can be obtained from the remaining dipoles in the same way.

2.3 Clustering

(1) Number the *k* dipoles that generate the subsets *A* and *B* in ascending order of their values and each number is called the number of the dipole or the two samples in the dipole. The *k* dipoles form *k* columns, which divides each of the subsets *A* and *B* into *k* clusters.

Definition 1 The two clusters, which belong to subsets *A* and *B* respectively, are called the corresponding cluster if they constitute a column of the set $A \bigcup B$.

For example, if *n*=10 then *k*=5. One possible result is shown in Table 1, where each of the subsets A and B is partitioned into 5 clusters. The cluster $\{4\}$ of subset *A* and cluster $\{2\}$ of subset *B* constitute the second column, thus forming a corresponding cluster. There are five corresponding clusters in Table 1.

Tab.1. When k = 5

A	Sample	1	4	6	8	10
	Numbering	(1)	(2)	(3)	(4)	(5)
B	Sample	3	2	5	7	6
	Numbering	(1)	(2)	(3)	(4)	(5)

Tab.2. When k = 4

A	Numbering	(1)	(2)	(3)	(4)	(5)
		(3)		(1)		
B	Numbering	(1)	(2)	(3)	(4)	(5)
			(5)			(2)

Definition 2 The consistency criterion $\eta_c = (p - \Delta k)/p$, where p is the total number of samples, Δk is the total number of those columns in which the numberings of samples of corresponding cluster is the same. (Identical columns should be counted repeatedly).

In Table 1, Δk =5 and $\eta_c = (5-5)/5 = 0$.

(2) Divide the subsets A and B into $k-1$ clusters. Cluster two of the closest samples of the set *A* into one cluster, and do the same with the set *B*. For example, the closest samples (1, 3) in *A* and the closest samples (2, 5) in *B* are clustered into one clustered into one cluster respectively. Then subsets *A* and *B* are divided into $k-1$clusters respectively, still the set $A \bigcup B$is partitioned into *k* columns. The two clusters, which lie in the same column and belong to A and B respectively, are also called a pair of corresponding clusters.

In this case, only a corresponding cluster in the fourth column has the same numbering of the samples (see Table 2). So Δk =1 and $\eta_c = (5-1)/5 = 0.8$.

Notice that the figures in Table 2 are the numberings of all samples rather than the samples themselves.

(3) Partition the subsets A and B into $k-2$ clusters in the same way. The distance between two clusters is determined by the distance between closest samples from the two clusters. The set $A \bigcup B$ is still partitioned into *k* columns.

Continue this process until the subsets *A* and *B* are clustered into two clusters respectively.

(4) When $\eta_c = 0$, the numbering of all corresponding clusters is identical, hence each pair of corresponding clusters can be merged into a cluster among the cluster candidates.

2.4 Find out the unique optimal clustering scheme using testing sets C and D

We do clustering again on testing sets *C* and *D* with the same way as that used in clustering process on the sets *A* and *B*. Then we observe the clustering schemes with $\eta_c = 0$ on sets *A* and *B*. If the clustering scheme with the same clustering number also satisfies $\eta_c = 0$ on sets *C* and *D*, then this one is the optimal clustering scheme we are looking for.

Remark 1 At the beginning of the clustering process, the value of the first scheme criterion η_c is always zero. However this scheme cannot be taken as the optimal clustering candidate since each

cluster is now only composed of a single dipole.

Remark 2 When $\eta_c \neq 0$, we are not sure that the smaller the value of η_c, the better the effect of the clustering. Thus no optimal clustering scheme can be determined.

3 GMDH cluster analysis method

We know from its algorithm steps in the Section 1.3 that OCA can automatically and objectively determine the clustering number and find out the optimal clustering scheme. However OCA has its limitations in practice.

3.1 The shortcomings of OCA

When applying OCA, the following two situations may occur.

(1) No clustering scheme with $\eta_c = 0$ occurs on set $A \bigcup B$ except the case of $k = \lfloor n/2 \rfloor$. Thus no optimal clustering scheme can be determined when applying OCA to this situation (see Section 3.1 and Remark 2 of Section 1.4).

(2) More than one clustering schemes with $\eta_c = 0$ may occur on set $A \bigcup B$ except the case of $k = \lfloor n/2 \rfloor$, and no clustering scheme with $\eta_c = 0$ occurs on set $C \bigcup D$. Thus we cannot objectively determine the unique optimal clustering scheme (see Section 3.2).

3.2 The new consistency criterion

Can we find out a more effective criterion than the original consistency criterion η_c? This means to those clustering scheme with $\eta_c = 0$, the new criterion also reaches its optimal value (see the proposition below), and when no clustering scheme with $\eta_c = 0$ can be found (i.e. no optimal clustering scheme can be determined by the original criterion), the optimal clustering scheme can still be determined by the new criterion.

Suppose that the sets A and B are clustered into *k* clusters respectively (Identical clusters should not be counted repeatedly.), and the total dipoles numbering is *q*.

Definition 3 Check the columns composed by set $A \bigcup B$. If the number of dipoles in the same column have two or more, such dipoles are called *determined dipoles*. The total number of such dipoles in the set $A \bigcup B$ is denoted by *p*. The clusters produced by such column is called *determined cluster*s and the total number of these clusters is denoted by *r* (Identical clusters should not be counted repeatedly, and the counting is done only in A or in B).

The total number of undetermined dipoles is denoted *M*, where $M = q - p$, and the total number of undetermined clusters is denoted *N*, where $N = k - r$. Clearly we have $M \geq N$, so we give the following definition of the new consistency criterion.

Definition 4 The new criterion is denoted as $\theta_c = M - N$, where θ_c is a non-negative integer which is the difference of the number of undetermined dipoles and the number of undetermined clustersClearly the smaller the difference, the higher the degree of consistency of corresponding clusters in the set $A \bigcup B$. Thus if we consider the columns generated from set $A \bigcup B$ as resulting clusters, then the smaller θ_c is, the more optimal the clustering scheme is. The following proposition explains the relationship between η_c and θ_c.

Proposition In the same clustering scheme, $\eta_c = 0$ implies $\theta_c = 0$.

Proof Suppose $\eta_c = 0$. Then, by definition 2, it means the sample numbering in two clusters of the corresponding clusters in each column is identical. We discuss it in two cases of the columns of set $A \bigcup B$. (1) If the number of dipoles in a column is greater than 1, then, by Definition 3, the clusters and dipoles in this column are already determined, and the values of *M* and *N* will not be affected in this case. (2) If the number of dipoles in a column is 1, then, by definition 3, both the cluster and dipole in this column are undetermined and the values of both *M* and *N* should be increased by 1 respectively . In both cases, we have $M = N$, hence $\theta_c = 0$. This completes the proof.

Now we illustrate, by examples, the meaning of the new criterion θ_c. We first consider the example with $k = 5$ in Section 3.2.

A	（1,2）	（2,1）	（3）	（4）	（5,9）	（6）	（7）	（8）	（9,5）
	（7）	（7）		（5,9）	（4）		（1,2）		（4）
B	（1,7）	（2,1）	（3）	（4）	（5）	（6）	（7,1）	（8,2）	（9）
	（8,2）	（8,7）		(9)			（8,2）	（1,7）	（4）

The number of different clusters in subsets *A* and *B* are both $k = 5$， the total numbering *q* of dipoles is 9. Set $A \bigcup B$ is composed of 9 columns, in which the first , the second and the seventh columns have the same composition: each of these columns has the same dipoles $\{\begin{pmatrix}1\\1\end{pmatrix},\begin{pmatrix}2\\2\end{pmatrix},\begin{pmatrix}7\\7\end{pmatrix}\}$. The fourth and the ninth columns also have the same dipoles $\{\begin{pmatrix}4\\4\end{pmatrix},\begin{pmatrix}9\\9\end{pmatrix}\}$. Thus we have two determined cluster (r=2 and $N = k - r = 3$), and 5 determined dipoles $\{\{1,2,4,7,9\}\}$($p = 5$ and $M = q - p = 4$). Thus the new criterion $\theta_c = M - N = 1$. There are 4 undetermined dipoles $\{\begin{pmatrix}3\\3\end{pmatrix},\begin{pmatrix}5\\5\end{pmatrix},\begin{pmatrix}6\\6\end{pmatrix},\begin{pmatrix}8\\8\end{pmatrix}\}$, and 3 undetermined clusters.

Now we consider the clustering scheme with *k* = 4 in Section 3.2.

A	（1,2）	（2,1）	（3）	（4）	（5,9）	（6）	（7,8）	（7,8）	（9,5）
	（7,8）	（7,8）		（5,9）	（4）		（1,2）	（1,2）	（4）
B	（1,7）	（2,1）	（3）	（4）	（5）	（6）	（7,1）	（8,2）	（9）
	（8,2）	（8,7）		(5,9)	(4,9)		（8,2）	（1,7）	（4,5）

We analyze the clustering process from $k=5$ to $k=4$. Actually（8）and（1,2,7）are merged in set A while（5）and（9,4）are merged in set B, thus reducing the number of the undetermined dipoles and so the value of θ_c. Now, the sample numbering of the two clusters in each corresponding cluster of set $A \bigcup B$ is completely identical, so $\eta_c = 0$. Thus we still have $\theta_c = 0$ by Definition 4.

4 The demonstration comparison between OCA and GMDH clustering methods

In the example of section 3.1, the optimal clustering scheme cannot be found using OCA. In the example of section 3.2, OCA cannot determine the unique optimal clustering scheme. However, using GCA, we can determine the optimal scheme in both cases.

4.1 Regional cluster analysis based on germination of certain tree seeds

The analysis problem was reported in [7]. We measure the average sprouting rate（X_1）and germinating energy（X_2）of certain tree in 12 different regions as shown in Table 3.

Tab.3. the germination conditions of certain tree seeds in different regions

Area code	X_1	X_2	Area code	X_1	X_2
1	0.707	0.385	7	0.877	0.713
2	0.600	0.433	8	0.513	0.353
3	0.693	0.505	9	0.815	0.675
4	0.717	0.343	10	0.633	0.465
5	0.688	0.605	11	0.740	0.580
6	0.533	0.380	12	0.777	0.723

We use both OCA and GCA to conduct cluster analysis. The main process are as follows.

(1) Generating dipoles and sample subsets A and B

Using the method in Section 1.2, we have dipoles $\begin{pmatrix}8\\6\end{pmatrix},\begin{pmatrix}4\\1\end{pmatrix},\begin{pmatrix}2\\10\end{pmatrix},\begin{pmatrix}12\\9\end{pmatrix}$ *and* $\begin{pmatrix}11\\5\end{pmatrix}$, thus obtaining sample subsets as follows:

A: 8, 4, 2, 12, 11

B: 6, 1, 10, 9, 5

(2) Obtaining the value of each clustering scheme（The clustering process is omitted)

$$k=5 \quad \eta_c = (5-5)/5 = 0, \qquad \theta_c = 5-5 = 0$$

$$k=4 \quad \eta_c = (5-2)/5 = 0.6, \qquad \theta_c = 5-4 = 1$$

$$k=3 \quad \eta_c = (5-0)/5 = 1, \qquad \theta_c = 5-3 = 2$$

$$k=2 \quad \eta_c = (5-0)/5 = 1, \qquad \theta_c = 2-1 = 1$$

From the clustering process we know that, using OCA, we have $\eta_c \neq 0$, except for $\eta_c = 0$ when $k = 5$ at the beginning. Thus we cannot find the optimal clustering scheme with consistency criterion η_c in this example. To the new consistency criterion θ_c. the smaller its value is, the better the clustering scheme is. Computing with this new criterion, we know that "the optimal candidate scheme" of the cluster occurs in the cases of $k = 4$ and $k = 2$. The optimal clustering schemes selected from sets *C* and *D* are divided into two clusters (The results are the same as those in [7] and the process is omitted.):

The first cluster is {8，6，2，10，1，4，3}；the second cluster is {12，9，11，7，5}.

4.2 Cluster analysis of nations（regions）based on information infrastructure

We conduct clustering of 20 countries(regions) according to their development of information infrastructure[8]. There are six development attributes（variables）of information infrastructure: (1)CALL: the length of telephone line of one thousand persons; (2) MOVECALL : the number of cellular mobile phones owned by one thousand families; (3) FEE: international telephone cost of every 3 minutes in peak-hour, (4) COMPUTER: the number of computers owned by one thousand persons; (5) MIPS: million instructions per second; (6) NET: internet householders of one thousand persons. The development status of information infrastructure in the 20 countries (regions) are shown in Table 4[8].

(1) Generating dipoles and subsets *A* and *B*

$$\text{Dipoles: } \binom{12}{10}, \binom{14}{13}, \binom{9}{8}, \binom{19}{3}, \binom{20}{7}, \binom{6}{4}, \binom{18}{17}, \binom{16}{15}, \binom{5}{2}$$

Samples in set *A*: 12, 14, 9, 19, 20, 6, 18, 16, 5

Samples in set *B*: 10, 13, 8, 3, 7, 4, 17, 15, 2

(2) Clustering with consistency criterion η_c on subsets *A* and *B*

We know from Table 5 that (except for *k* = 9) $\eta_c = 0$ when *k* = 4, *k* = 3 and *k* = 2. We then have to determine which one is the optimal clustering scheme.

(3) Clustering with consistency criterion η_c on subsets *C* and *D*

The produced dipoles of sets C and D are as follows：

Tab.4. The development of information infrastructure of these nations (regions) [9]

Serial number	Countries (Regions)	CALL	MOVECALL	FEE	COMPUTER	MIPS	NET
1	America	631.60	161.90	0.36	403.00	26073	35.34
2	Japanese	498.40	143.20	3.57	176.00	10223	6.26
3	Germany	557.6	70.60	2.18	199.00	11571	9.48
4	Sweden	684.10	281.80	1.40	286.00	16660	29.39
5	Switzerland	644.00	93.50	1.98	234.00	13621	22.68
6	Denmark	620.30	248.40	2.56	296.00	17210	21.84
7	Singapore	498.40	147.5	2.50	284.00	13578	13.49
8	Taiwan of China	469.40	56.10	3.68	119.00	6911	1.72
9	Korean	434.5	73.00	3.36	99.00	5759	1.66
10	Brazil	81.90	16.30	3.02	19.00	876	0.52
11	Chile	138.60	8.20	1.40	31.00	1411	1.28
12	Mexico	92.20	9.80	2.61	31.00	1751	0.35
13	Russia	174.90	5.00	5.12	24.00	1101	0.48
14	Poland	169.00	6.50	3.68	40.00	1796	1.45
15	Hungary	262.20	49.40	2.66	68.00	3067	3.09
16	Malaysia	195.50	88.40	4.19	53.00	2734	1.25
17	Thailand	78.60	27.80	4.95	22.00	1662	0.11
18	India	13.60	0.30	6.28	2.00	101	0.01
19	French	559.10	42.90	1.27	201.00	11702	4.76
20	Britain	521.10	122.50	0.98	248.00	14461	11.91

We apply OCA on *C* and *D*. Then, except the case of $k = 9$， there is no clustering scheme with $\eta_c = 0$ appears (see Table 5). Thus no additional information is available for determining the unique optimal clustering scheme, and the consistency criterion η_c is ineffective.

Tab.5. The value of consistency criterion η_c in each clustering scheme

	On sets *A*, *B*	On sets *C*, *D*
K = 9	η_c =0	η_c =0
K=8	η_c =0.333	η_c =0.333
K=7	η_c =0.556	η_c =0.778
K=6	η_c =0.667	η_c =1
K=5	η_c =0.778	η_c =1
K=4	η_c =0，	η_c =1
K=3	η_c =0，	η_c =1
K=2	η_c ==0	η_c =1

(4) Clustering with the new consistency criterion θ_c

We repeat steps (2) and (3) in this section, but cluster with the new consistency criterion θ_c. The results are shown in Table 6. We have three candidates of the optimal clustering on sets *A* and *B*: *k* = 4，*k* =3 and *k* = 2. Then we use sets *C* and *D* to detect these candidate schemes. We have θ_c = 3 as *k* = 4, θ_c = 2 as *k* = 3 and θ_c = 1 as *k* = 2. θ_c reaches its minimum as *k* = 2, so the optimal clustering

scheme is divided into two clusters.

The first cluster is {10，12，13，14，15，16，17，18，11}. They represent Brazil, Mexico, Russia，Poland, Hungary, Malaysia, Thailand, Chile and India. They are nations in transformation or developing countries in Asian and Latin America. They have less developed economy, weak infrastructure and are insufficient in information infrastructure.

The second cluster is {8，9，3，19，20，7，6，4，5，2，1}. They represent Taiwan of China, South Korea, Germany, France, Britain, Singapore, Denmark, Sweden, Switzerland, Japan and United States. These countries(regions) have highly developed information infrastructure.

Tab.6. The value of consistency criterion θ_c in each clustering scheme

	On sets *A*, *B*	On sets *C*, *D*
K=9	θ_c =9-9=0	θ_c =9-9=0
K=8	θ_c =9-8=1	θ_c =9-8=1
K=7	θ_c =7-6=1	θ_c =9-7=2
.K=6	θ_c =6-5=1	θ_c =9-6=3
K=5	θ_c =6-5=1	θ_c =7-4=3
K=4	θ_c =2-2=0	θ_c =6-3=3
K=3	θ_c =1-1=0	θ_c =3-1=2
K=2	θ_c =0-0=0	θ_c =1-0=1

5 Conclusions

Compared with the usual statistical clustering method, OCA can automatically and objectively determine the number of clusters and find out the optimal clustering scheme. However in many cases OCA is unable to determine the optimal scheme due to the shortcomings of its consistency criterion (which is the core component of OCA). In order to solve the problem, we proposed a new consistency criterion. The proposed GMDH clustering method based on the new consistency criterion extends the application scope of clustering objects. Our theoretical analysis and practical examples demonstrate that the new consistency criterion is not only more powerful than the original consistency criteria, but also has definite practical significance.

References

[1] Jiawei Han, Micheline Kamber.(2001). Data Mining: Concepts and Techniques[M]. Morgan Kaufmann Publishers, Inc.

[2] Mehra R.K. Group method of data handling (GMDH): review and experience[C]. in Proceedings of the IEEE Conference on Decision and Control，1977,29-34.

[3] Ivakhnenko A G，Mueller J A.(1992). Parametric and nonparametric selection procedures in experimental systems analysis[J]. SAMS, 9(5): 157-175.

[4] Ivakhnenko A G. at all. Objective selection of optimal clusterization of a data sample during compensation of non-robust random interference. Journal of Automation and Information Sciences[J] 1993,26(3):45-56.

[5] Mueller J A, Lemke F.(2000). Self-organising data mining[M]. Dresden, Berlin: Libri Books.

[6] Madala H R, Ivakhnenko A G.(1994). Inductive learning algorithms for complex systems modeling[M]. Boca Raton,London,Tokyo：CRC Press.Inc.

[7] Wenshuang Sun, Lanxiang Chen. Multivariate Statistical Analysis[M].Beijing: Higher Education Press. 1994.

[8] Xiulin Yu, Xuesong Ren. Multivariate Statistical Analysis[M].Beijing: China Statistics Press. 1999.