

Criterion of Congruence as a Criterion of Selection for Clusterization

Natalya Ivakhnenko

*International Research and Training Center of Information Technologies and Systems
of the NAS and MES of Ukraine, Glushkov ave, 40, Kyiv-680, MSP 03680, Ukraine
Phone: +38(044) 502-63-37*

dep175@irtc.org.ua

Abstract. *The criterion of congruence consists in the comparison of two or more clusterizations on two square arrays of specially organized points. Such arrays are named “faces” of a given clusterisation. These faces help to compare arrays with various quantity of clusters and various number of points in them simultaneously, and this is because the criterion is called as a congruent one. As in the case of the algorithms of self-organization, this is used firstly for finding the multitudes of arguments for two arrays, selecting better ones for electing the criterion. Then, knowing a few better multitudes of arguments, one find, using already known procedures in the first part, the multitudes of arguments for the full array. The proposed criterion would open a set of other selection ones in the future.*

Keywords

Inductive modelling, GMDH, criterion of congruence, clusterization, self-organization.

1 Introduction

Let us talk a little about the theory of Self-organization [1-4]. It was created by academic Alexey G. Ivakhnenko in the Ukrainian Academy of Sciences. The first work appeared about forty years ago. During this time it has been spreading more and more widely.

We just note that this theory does not compete with ordinary mathematics because it occupies its own definite place – “niche” in the theory of clusterization, modeling and forecasting of processes. This theory may find the model of the signal when its noise is two or three times more than signal itself in the noised experimental data. It may also generate the information about the model of linear process when it is described by five, six and more values, having only four or five points of this process. If your data samples don't have noise or they have very little noise you can use the methods of the ordinary mathematics you know very well.

Let us consider the main statements of this theory and look at one of its algorithms with the new criterion of selection. Almost all the algorithms of the theory of self-organization are based on four main statements (positions). We may imagine them like four whales holding it on their backs.

First statement

It is necessary for the sorting out to be present. Without the sorting out there is no selection, and without the selection there is no task! You may sort out the values, signals, parameters, signs, frequencies, coordinates, their changes, their combinations and so on.

Second statement

It is necessary for the external supplement to be present. Most often it consists of the construction of two or more subsets. One of them helps us to find the coefficients of model and the rest of them to control our solutions with the calculation of selection criterion.

Third statement

Certainly, it is necessary for the criterion itself to be present. There are many criteria and their combinations. Almost every new algorithm has new criterion of selection.

Fourth statement

.Every algorithm must have the freedom of choice F . It is the number of the best solutions. If you have a “ little ” task (only six or seven variables) you may use the combinatorial algorithm. This is the algorithm of full sorting out of variables. But even in this case you must propose to your customer F best models despite the fact that they have almost equal criterion. According to his wishes or notions about process he will choose one of them. You will be able to see this statement in action when you have more than ten variables, and the combinatorial algorithm can not be used. Then you must solve the task with the Multilayered Iterative algorithms. Here there are F best solutions passing from each layer of selection to next one. You may read about it more detail in the books about this method [1-4]. For the introduction of our selection criterion let us consider some famous simple concepts.

2 Congruence

As you know from an elementary geometry congruent figures are those which have two or more equal parameters. So, the triangles are congruent if their sides and angles are equal, though they may have different color, weight or height.

3 Dipoles and their parameters

It is necessary to remind also that dipole is a subset of any two points. These points are named its poles, and the distance between them its size.

4 Division of the experimental data

Let us consider the array of data sample $x[M,N]$, where M - the quantity of values and N - the quantity of points. Standartize this array. Let it will be the array $X[M,N]$ in which each column is value from zero to unity. Now consider any intermediate ensemble of m values (where m - the quantity of values in this ensemble). It is presented as the array $X[m,N]$. For our example, let N equal twelve ($N=12$). First, let us divide this array $X[m,12]$ in two such subsets that analogues points of the array $XB[m,6]$ would be like the array $XA[m,6]$. There are many methods of such operation.

For example, here is one of them. Place all twelve points of the array $X[m,12]$ along the line of the between-point distance from the beginning of the coordinates. Then place all the even points of this line into the array $XA[m,6]$, and place all the odd points into the array $XB[m,6]$.

5 The “ tree” of the clusterization

Now we consider one of these arrays. Let us take the array $XA[m,6]$ and construct the “tree” of clusterization. This array has six points ($n=6$). Designate them as 1,2,3,4,5,6. Constructing dipoles from them, find distances between these points, and sort these distances according to minimum of their values. Let such array of poles be as following :

2 1 1 5
5 3 4 6 . . .

For our example the number of these dipoles will be $n*(n-1)/2=6*5/2=15$. This number is the sum of numbers from unit to n , as the sum of elements of the triangular matrix of between-point distances.

Let us introduce the vector $IR(6)$ that will be filled according to the addresses of poles and steps of clusterization “tree” construction. Here we have meant of the method of ordinary clusterization.

6 The construction of congruence criterion

This criterion consists of the comparison of both clusterization on two especially organized square arrays of points. Each of these squares is named “face” (Russian – Litzo) of given clusterisation. During the construction of our criterion

we should build the array $IA[6,6]$ and the array $IB[6,6]$. We shall call them “Faces” for the array $XA[m,6]$ and the array $XB[m,6]$ accordingly. Create one of them.

Let it the array $IA[6,6]$. According to following simple rules take again a vector $IR[6]$, having before hand nulled it and the array $IA[6,6]$. It is very important to null them every time before the construction of “face” for each new intermediate ensemble. Then fill the diagonal elements of this array from one to $N/2$ (for our example, $n=N/2=6$). By analogy with the construction of tree – clusterisation we consider the reorganized dipole row in the order of size (from minimum to maximum size). We may also receive these dipoles from the triangular-matrix of the between- point distances, because it is the list of these distances, and their addresses of location on this matrix are the addresses of the dipole’s poles. For better presentation let this dipoles’ row for the array $XA[m,6]$ will be such as we had above during the construction of “tree” :

2 1 1 5
5 3 4 6

Then gradually considering these dipoles write the values (according to their addresses) in vector $IR[6]$

0 2 0 0 5 0

and in the array $IA[6,6]$ simultaneously during the transfer from dipole to next one. Note that both of poles must be equal in their “rights” because in the array $IA[6,6]$ we write dipole twice. We may see the filling of the array $IA[6,6]$ at next figure :

1	0	0	0	0	0	1
0	2	0	0	5	0	2, 5
0	0	3	0	0	0	or 3
0	0	0	4	0	0	4
0	2	0	0	5	0	2, 5
0	0	0	0	0	6	6

Our first step will be such: as we may see in this figure 2 and 5 took places corresponding to their addresses. Note that for first dipole we did not offend rights of each pole. We write 2 and 5 both in the second and the fifth lines. It is important point because we will also follow the same operation in the future. Let us repeat this operation with the rest of dipoles of list 2 until the vector $IR[6]$ will be complete. Let us image the picture of this moment, and the last step of clusterisation for the face construction will be such :

1	0	3	4	0	0	1, 3, 4
0	2	0	0	5	0	2, 5
1	0	3	0	0	0	or 1, 3
1	0	0	4	0	0	1, 4
0	2	0	0	5	6	2, 5, 6
0	0	0	0	5	6	5, 6

Now draw the final face picture for the array $XA[m,6]$ adding some elements in the square array $IA[6,6]$. If any two lines have at least one common element they are rewritten so that all their elements become equal. Gradually passing from the first to last lines we will make square. This matrix will be face for the array $XA[m,6]$. Let’s draw this moment so :

1	0	3	4	0	0	1, 3, 4
0	2	0	0	5	6	2, 5, 6
1	0	3	4	0	0	or 1, 3, 4
1	0	3	4	0	0	1, 3, 4
0	2	0	0	5	6	2, 5, 6
0	2	0	0	5	6	2, 5, 6

How we can see this picture looks like the list of six clusters. There are three identical clusters. It is easy to see that we received only two different clusters with next list of points:

first cluster : 1 3 4
 second cluster : 2 5 6

Repeat such procedure for the array $XB[m,6]$.
 For simplicity let its face will be the array $IB[6,6]$ which we shall draw such figure :
 It was constructed on the basis of the next row of dipoles:

$$\begin{array}{cccc} 2 & 1 & 1 & 4 \\ 4 & 3 & 5 & 6 \dots \end{array} \quad (3)$$

It differs from (2) in that 4 and 5 have changed places. Then “face” of the array $XB[m,6]$ will be such:

$$\begin{array}{cccccc} 1 & 0 & 3 & 0 & 5 & 0 & 1, 3, 5 \\ 0 & 2 & 0 & 4 & 0 & 6 & 2, 4, 6 \\ 1 & 0 & 3 & 0 & 5 & 0 & \text{or } 1, 3, 5 \\ 0 & 2 & 0 & 4 & 0 & 6 & 2, 4, 6 \\ 1 & 0 & 3 & 0 & 5 & 0 & 1, 3, 5 \\ 0 & 2 & 0 & 4 & 0 & 6 & 2, 4, 6 \end{array}$$

Then let us put these faces one on top of the other and calculate the number of unequal corresponding elements.

$$\begin{array}{cccccc} 0 & 0 & 0 & \$ & \$ & 0 \\ 0 & 0 & 0 & \$ & \$ & 0 \\ 0 & 0 & 0 & \$ & \$ & 0 \\ \$ & \$ & \$ & 0 & 0 & \$ \\ \$ & \$ & \$ & 0 & 0 & \$ \\ 0 & 0 & 0 & \$ & \$ & 0 \end{array}$$

Their number will be sixteen. This is the value of the congruence criterion. It gives us the possibility to evaluate the clusterization both by quality (the numbers of clusters) and by quality (the qualitative composition of clusters) with ONE figure. Precisely for this reason this criterion differs from all earlier used. It is the first criterion among similar ones which (we hope) will appear in the near future. We call it the congruence criterion. Besides we should note that this criterion is an example of “ the principle of unfinal decisions”. You can see that calculation of criterion is done before the final decision about the clusterization of given array’s points. Maybe some of you will say that there are many such face - squares, which have equal criterions because we consider it as a “sum” of two concepts. But you must remember that we have picked up beforehand two arrays that had similar points. These points were situated on close enough distances. In other words we work in narrow enough range of this criterion.

7 Comment to the sorting out

We may note that the division of two subsets must take place for *each intermediate ensemble of the variables*. It is very important for results. All earlier algorithms made the division before the sorting out. It is and was a great mistake because it was not done in the space of M variables. Remember that M is full quality of variables in the input of algorithm. So, the division of two subsets in the M -space for the construction of research model may be wrong and the time of model search may significantly increase. For example let us consider five points on the axis x_j . Let they be at equal distance between each other and have next order 1, 2, 3, 4, 5. Let on other axis these points may lay in another order. So during the division of points on the subsets most often different points will be in other order. As a rule when we change the size of space we exchange the order of points.

We hope that proposed criterion will open new area of other selection criterions in the future.

References

- [1] Ivakhnenko A.G., Stepashko V.S. *Pomekhoustoychivost' modelirovaniya (Noise – immunity of modeling)*. Naukova Dumka, Kiev, 1985.
- [2] Ivakhnenko A.G., Yurachkovskiy. *Modelirovaniye slzhnykh system po eksperimentalnym dannym (“Modeling of Complex Systems for Experimental Data)*. Radio I svyaz, Moscow, 1987.
- [3] Vorontsov K.V. *About combinatorial exit to the learning algorithms*. Reception for the eleventh United conference in Putscino, Russian, 2003.