

# Parallel GMDH algorithm with successive selection of informative arguments for effective solving high-dimensional modelling problems

Volodymyr Stepashko, Serhiy Yefimenko, Oleksandr Samoilenko

*International Research and Training Center of Information Technologies and Systems of NAS of Ukraine, prospect Akademika Glushkova, 40, Kyiv, 03680, Ukraine*

stepashko@irtc.org.ua syefim@ukr.net soa\_pga@mail.ru

**Abstract.** *In recent investigations we considered GMDH algorithms for solving problems with a large number of arguments based on successive selection of the most informative arguments. These algorithms build models very quickly but the accuracy of obtained models is often not very high. The models quality being built in such a way depends on the quality of informative arguments selection. To increase this quality in the algorithm with successive selection, the inverted structures are used. As a result we have the quality enhancement but the speed of the model building is somewhat reduced. To solve this problem the parallel algorithm is implemented. Thus the main goals of our work are improvement of the method with successive selection of arguments and the parallel algorithm implementing in this method for the enhancement of the quality and effectiveness of the informative arguments definition.*

*This paper considers the main aspects of the algorithm and results of its performing. The test experiments for the parallel implementation of algorithm with successive selection of arguments using inverted structures on a cluster system are carried out. The results of these experiments confirm effectiveness of the method.*

## Keywords

Inductive modelling, combinatorial GMDH algorithm, parallel computing, cluster system, successive selection, informative arguments,

## 1 Introduction

Inductive GMDH algorithms suppose examination of all possible variants of task solving and selection of the best variant (model). We include to the sample as more arguments as possible to build the most exact model. However if number of arguments is very great then examination of all variants takes too much time and is often impossible. Therefore, there is the problem of acceleration of the combinatorial algorithm namely by means of informative arguments selection. The problem has been investigated in [1], [2] where it was suggested to estimate the level of arguments informativity regarding to the module of the argument correlation coefficient with the output variable (MCC). The algorithm of the argument informativity definition [3] is on the bases of our investigation. To improve the effectiveness of this algorithm we propose to use a method with successive selection of arguments and its parallel implementation.

## 2 Problem statement

The problem of structural identification may be considered as determination of the best in certain sense separation of input variables  $X$  into informative and noninformative ones.

So, as a whole the problem of identification consists in forming by sampling data  $W = (X : y)$  a certain set  $\mathfrak{S}$  of models of different structure of the form  $\hat{y}_f = f(X, \hat{\theta}_f)$  and search of the optimal model by minimum of the given criterion  $CR(\cdot)$

$$f^* = \arg \min_{f \in \mathfrak{S}} CR(y, f(X, \hat{\theta}_f)), \quad (1)$$

where estimates of parameters  $\hat{\theta}_f$  for every  $f \in \mathfrak{S}$  is solution of one more problem as

$$\hat{\theta}_f = \arg \min_{\theta_j \in R^{s_j}} Q(y, X, \theta_f) \quad (2)$$

in which  $Q(\cdot) \neq CR(\cdot)$  is performance criterion of solution of the problem of parametric identification of every particular model, generated in the problem of structural identification (2), and  $s_f$  is the number of parameters being estimated.

For definition of model structure we will use the structural vector that defines which exactly vectors from matrix  $X$  will be included in the model. We call  $m$ -dimensional vector  $d = \{d_1, \dots, d_m\}$ , which consists of  $s$  units and  $m - s$  zeroes,  $s = \overline{0, m}$ , which indicate the presence (for  $s = \overline{0, m}$ ) or the absence (for  $d_j = 0$ ) of the corresponding component of the input vector  $x$  in the verified model, as the structural vector. Here the number  $s$  is called complexity of a model.

Then structure of model  $f$  is a matrix  $X_d$  consisting of vectors  $x_j$ ,  $j = 1, \dots, s$  from matrix  $X$  defining by structural vector  $d$ .

Every model  $f$  in  $\mathfrak{S}$  differs from other models of this set by its structure. That is why the number of all models of  $\mathfrak{S}$  is equal to number of structural vectors, which could be built for  $m$  input variables.

The number of  $s$  arrangement units variants in  $m$ -dimensional binary vector is equal to  $C_m^s = \frac{m!}{s!(m-s)!}$  therefore, the number of all structural vectors is calculated by formula

$$P_m = \sum_{s=1}^m C_m^s = 2^m - 1. \quad (3)$$

When the arguments number is greater than 30, the exhaustive search for the acceptable limit of time is often impossible.

### 3 Solving of the problem

Let us start with an example:  $m=20$ ,  $n=50$ ,  $s_0=10$ , and analyze the dependence of the criterion  $AR$  on the model complexity  $s$ .

As it is evident from (3), removing of an argument from the set arguments halves the time of searching. Consequently, for the acceleration of finding of the best model it is needed to find such arguments which will not substantially influence on the model and to remove them from the set, leaving most informing. Such approach is offered in [1], where it is suggested to estimate the level of arguments informativeness regarding to the module of the argument correlation coefficient with the output variable. In [3] it is shown that the level of arguments informativeness can be estimated considering how many of the best models contain this argument.

On the basis of these results, let us consider the algorithm of successive selection of informative arguments. Models are selected by this algorithm using the combinatorial algorithm COMBI (discovered by V. S. Stepashko) with successive complication. The complexity of building models is increasing until the calculation time is permissible. The best models and arguments included to these models are examined then only. A new set which consists only of those arguments taking active part in forming the best models is thus formed. Further models are built on this new set and the sequence of such operations are again repeated until the set will contain so many arguments that it would be possible to perform an exhaustive search.

We describe algorithm of successive selection of informative arguments (Alg. 1) as the following steps:

*step 1*: using formula (3) calculate  $s_{\max}$ , where  $s_{\max}$  defines the maximum of all models complexity we can build during the given time limit;

*step 2*: build all models of complexity  $s = \overline{1 \dots s_{\max}}$  using COMBI algorithm with successive complication;

*step 3*: select a subset of  $F$  the best models by an external criterion;

*step 4*: rank all the arguments being contained in this  $F$  models by the coefficient  $q_i, i=1 \dots m$ , specifying the frequency of an  $i$ -th argument occurrence in the best models;

*step 5*: form a new sample by removing the arguments with the least values of  $q_i$ ;

*step 6*: perform the exhaustive search of models if the amount of arguments in the new sample is acceptable or return to the step 1 otherwise.

Let us investigate the effectiveness of this algorithm at first theoretically.

When the task with  $m=20, s_0=10$  is solved by COMBI with successive complication, the amount  $P_m$  of all possible models containing no more than  $s_0$  arguments is calculated with the use of formula (3).

$$P_m = \sum_{j=1}^{10} C_{20}^j = 616665 \tag{4}$$

The models amount built by the Alg. 1 to get the result of the exhaustive search with 20 arguments is equal to 76842 that is considerably less than it was by the COMBI (616665, see (4)). As for the computing time, the figures are as follows: 3 sec for Alg. 1 and 24 sec for COMBI. The combinatorial algorithm with the exhaustive search finds the same model in 48 sec.

Let us consider the same task for  $m = 200$  using the Alg. 1.

**Tab. 1. Amount of models on each stage of the algorithm with successive selection of informative arguments,  $m=200$**

Stages	m	$s_{\max}$	$P_m$
Stage 1	200	2	20 100
Stage 2	51	3	22 151
Stage 3	50	3	20 875
Stage 4	25	5	68 405
Exhaustive search	4	4	15
Total amount of models, Alg. 1			63 141

This algorithm takes 4 sec for solving the problem. It's very quickly for this amount of arguments. But the accuracy of the built models is not sufficient.

To raise the quality of arguments extraction we propose to use the algorithm with inverted structures (Alg. 2). This algorithm is based on algorithm with successive selection of informative arguments, but with an addition. On every stage we will consider the models of complexity  $s$  together with the "conjugate" models of complexity  $m-s$  notably the models with invert structures. For sample invert structure of 10100 is a 01011 structure etc.

We describe algorithm of successive selection with inverted structures (Alg. 2) as the following steps:

*step 1*: using formula (3) calculate  $s_{\max}$ , where  $s_{\max}$  defines the maximum of all models complexity we can build during the given time limit;

*step 2*: build all models of complexity  $s = \overline{1 \dots s_{\max}}$  using COMBI algorithm with successive complication;

*step 3*: select a subset of  $F$  the best models by an external criterion;

*step 4*: rank all the arguments being contained in this  $F$  models by the coefficient  $q_{Ii}, i=1 \dots m$ , specifying the frequency of an  $i$ -th argument occurrence in the best models;

*step 5*: build all models of complexity  $s = n - s_{\max} \dots n - 1$  (models with inverted structures) using COMBI algorithm with successive complication;

*step 6*: select a subset of  $F$  the best models by an external criterion;

*step 7*: rank all the arguments being contained in this  $F$  models by the coefficient  $q_{2i}$ ,  $i=1\dots m$ , specifying the frequency of an  $i$ -th argument occurrence in the best models;

*step 8*: form a new sample by removing the arguments with the least values of  $q_i$ , where  $q_i = \frac{q_{1i} + q_{2i}}{2}$

*step 9*: perform the exhaustive search of models if the amount of arguments in the new sample is acceptable or return to the step 1 otherwise.

Let us consider the task for  $m = 200$  using this algorithm and compare it with Alg.1 (Tab. 2).

**Tab. 2. Comparison of the algorithms effectiveness,  $m=200$**

Algorithms	Amount of models	Time (sec)	Complexity (s)	AR
Alg. 1	63 141	4	4	31.97
Alg. 2	905 806	126	12	1.74

This algorithm solves the problem with 200 arguments using 65 stages. As a result we get all real arguments in model and very high quality increase. But the speed of the models building is considerably reduced. In order to raise the algorithm speed we suppose to use the parallel method.

## 4 The results of experiments

To study effectiveness of the parallel implementation of algorithm with successive selection of arguments using inverted structures we carried out test experiments on solving high dimensional problem of structural and parametrical identification with the use of cluster system scit-3 [4].

The experiments were executed as follows: the design matrix  $X$  of size  $50 \times 150$  (150 records for 50 arguments) was generated. Vector  $y$  was formed as a linear combination of the first ten arguments so that the true model looked like  $y = -3x_1 - 3x_2 + 5x_3 - x_4 - x_5 + 3x_6 + x_7 - 2x_8 + x_9 + x_{10}$  with addition of noise. We used criterion of regularity for the best model selection.

Table 3 represents maximum complexity on each stage of the algorithm for separation of all steps on 24 processors and implementation on one processor. It was calculated such that maximum models being built by each processor on each stage do not exceed 300 000 for Alg.1. Amount of built models by Alg.2 was twice as much. As table shows maximum complexity and therefore total amounts of models is higher in case of using parallel implementation.

**Tab. 3. Maximum complexity on each stage**

Stages	Amount of selected arguments	Maximum complexity, used 1 processor, Alg.1, Alg.2	Maximum complexity, used 24 processors, Alg.1, Alg.2
1	50	4	6
2	45	4	6
3	41	4	6
4	37	5	7
5	34	5	7
...	...	...	...
15	15	15	15
Total amount of models, Alg.1		1168872	24039735
Total amount of models, Alg.2		2337744	48079470

Finally enumeration of greater amount of models allows finding a model with a higher amount of true arguments (tab. 4).

**Tab. 4. True arguments in the best models**

	True arguments in the best model
Uniprocessing, Alg.1	x1, x2, x3
Uniprocessing, Alg.2	x1, x2, x3, x6, x8
Multiprocessing, Alg.1	x5, x6, x7, x8, x9
Multiprocessing, Alg.2	x2, x3, x4, x6, x7, x8, x9
True model	x1, x2, x3, x4, x5, x6, x7, x8, x9, x10

Table 5 demonstrates the algorithm efficiency for providing of the even loading on all processors of the cluster system (main processor is responsible for interprocessor communication).

**Tab. 5. Run-time of Alg.2 for 24 processors**

Number of processor	Time, seconds
1, main	1089.6
2	1087.7
3	1088.2
...	
24	1088.1

## 5 Conclusion

The results of test experiments for the parallel implementation of algorithm with successive selection of arguments using inverted structures on a cluster system confirm effectiveness of the algorithm.

The use of algorithm enables to essentially accelerate the retrieval of the best subset of arguments and to solve problems with considerably larger number of arguments compared to the ordinary combinatorial GMDH algorithm with exhaustive search of arguments.

## References

- [1] Ivakhnenko A.G.: Ivakhnenko G.A., Savchenko E.A., and Wunsch D.: Problems of Further Development of GMDH Algorithms: Part 2. *Pattern Recognition and Image Analysis*, Vol. 12, № 1, 2002, p.6-18.
- [2] Ivahnenko A.G., Ivahnenko G.A., Savchenko E.A.: Conception of the successive algorithmic approaching (lowering) to the exact decision of interpolation tasks of artificial intelligence. *Cybernetics and computing engineering*, № 127, 2000, p.47-58. (In Russian)
- [3] Stepashko V.S., Koppa Y.V.: The Experience of application of the ASTRID system for the design of economic processes from statistical data. *Cybernetics and computing engineering*. - 1998. - V.117. – p. 24-31. (In Russian)
- [4] <https://icybcluster.org.ua/>