GMDH-based Forecasting for the Estimation of Medicines Efficiency

Nina Kondrashova[1], Andriy Pavlov[1]

[1] *International Research and Training Center of Information Technologies and Systems of National Academy of Sciences of Ukraine, 03680, Kiev, av. Glushkova, 40, Ukraine*

e-mail: Nkondrashova@ukr.net, Me_Ovechka@bigmir.net

Abstract. Influence of different initial set division methods on GMDH forecasting models accuracy is researched. Investigations are carried out within an application task in medical diagnostic sphere, namely in the task of medicines efficiency estimation. Three forecasting models construction schemes and four different set division types are investigated. Models are obtained by the two-stage division modeling algorithm and its modifications proposed. Several studies directions in forecasting models construction are analyzed. Criteria and results of numerical experiments are given. Analysis result showed that best forecasting models pair were obtained as sum of trend model and model of remains, utilizing quasi-optimal set division of the incomplete subset space with adaptive forecast scheme. The models pair satisfied the specified ±5,1% error range of the patients main group set with 0,04 significance level.

Keywords

Forecasting model, GMDH algorithm, set division, adaptive forecast, thiol-disulfide ratio (TDR).

1 Introduction

Influence of different methods of initial set division on GMDH forecasting models accuracy which used adaptive as well as nonadaptive build scheme [1] is researched in this paper. Investigations are carried out within an application task in medical diagnostic sphere, namely in the task of medicines and nonmedicamental remedies efficiency estimation for individual therapeutic agents selection. Initial data of each patient are presented by rows containing five values of thiol-disulfide ratio (TDR) as a time function called TDR-gram (TDR diagram). TDR is some fight degree characteristic of human immune system. Doctor can make conclusion about medicine effectiveness basing upon results of TDR-gram – results of patient's blood samples cultivation with them. It is suggested to build forecasting models to reduce TDR time observation from 24 hour down to 1 hour which can be used as decision support for individual medicine selection. Forecasting models use first observation hour TDR values which comprises first three measurement records and forecast next two time records: 3rd hour and 24th hour. Let we have randomly selected initial data for limited patients X set, $X \subset \Xi$, where power $|X|$ = 28, Ξ is the patients general totality set whose blood samples were cultivated with the given medicine.

Forecasting models were obtained using two-stage division modeling algorithm (TSDMA) since traditional single-stage division into learning and testing sets did not allow to get models with satisfactory accuracy [2].

2 Theoretical Part

Initial X set is divided into two subsets randomly: W, $|W|$ = 18 called active and D, $|D|$=10 called checking, X=$W \cup D$, $W \cap D$=Ø. Let keep the same denotation for X, W, D patients sets and corresponding X, W, D data matrices which comprise patients' testing results: $X \leftrightarrow \mathrm{X}$, $W \leftrightarrow \mathrm{W}$, $D \leftrightarrow \mathrm{D}$. Subset D is used only for models verification obtained using W subset. Initial data set is presented by $\mathrm{X} = (\mathbf{x}_0, \mathbf{x}_1, ..., \mathbf{x}_{m-1}) \in R^m$ matrix. Models pair (one for each forecasting TDR value) were obtained using GMDH and different division methods of $W = A \cup B \cup C$ set into A – learning, B – testing, C – examination, $A \cap B = \varnothing$, $B \cap C = \varnothing$, $A \cap C = \varnothing$. Two-stage division include such stages: at the first stage

(v = 1) active W set is divided into U training set and C examination set, $W = U \cup C$, $U \cap C = \varnothing$, then at the second stage (v=2) U training set is divided into learning and testing sets, i.e. $U = A \cup B$, $A \cap B = \varnothing$. Four set division types are researched in this paper: random, similar "by dispersion", dissimilar "by dispersion" and quasi-optimal. First three of them are traditional for GMDH but for finding quasi-optimal division [3] one should minimize norm:

$$p_v^* = \underset{\rho_\ell^2 \neq 0, \ell \in [1, L_v]}{\arg\min} \left\| \mathbf{X}_{1v_\ell}^T \mathbf{X}_{1v_\ell} - \rho_\ell^2 \mathbf{X}_{2v_\ell}^T \mathbf{X}_{2v_\ell} \right\|, \quad L_v = \sum_{n=n_{\min_v}}^{n_{\max_v}} C_{N_v}^n, \; v = \overline{1,2}$$

where p_v^* is the optimal rows composition of $\mathbf{X}_1$ matrix; L_v is the division variants maximal number of v-*th* stage; N_v is the elements number (power) of appropriate partitionable set of v-*th* stage; $n_{\max_v}$, $n_{\min_v}$ are the maximal and minimal elements number (power) of the first set, ($N_v - n_{\max_v}$) – of the second set. Rows partitionable sets variability (m parameter) which are used at the mentioned stages [4] is researched in this paper. Sets which use all data (arguments) in the dividing process are called complete information sets ($\dim \mathrm{X} = n \times m$), while the sets having $m_1 \leq m$ in this process are called incomplete information sets.

2.1 Problem statement

Let's formulate the models construction problem comparing theirs accuracy at the different parts of the $W = A \cup B \cup C$ set and the whole X set. Let introduce denotation: $\mathbf{y}_i = \mathbf{x}_{2+i}/\mathbf{x}_0$, $\mathbf{y}_i \in \Re^1$, i = 1,2 $\mathbf{z}_1 \overset{\Delta}{=} \mathbf{x}_0$, $\mathbf{z}_2 = \mathbf{x}_1/\mathbf{x}_0$, $\mathbf{z}_3 = \mathbf{x}_2/\mathbf{x}_0$, matrices $\mathbf{z}$ = ($\mathbf{z}_1$, $\mathbf{z}_2$, $\mathbf{z}_3$), $\mathbf{y}$ = ($\mathbf{y}_1$,$\mathbf{y}_2$), Z and Y are matrix sets of the corresponding matrices: $\mathbf{z} \in Z$, $\mathbf{y} \in Y$, $Z \in \Re^3$, $Y \in \Re^2$, where $\Re$ is the real number space. Let $\hat{\mathbf{y}}$ matrix be the estimation of the $\mathbf{y}$ matrix, $Q(\mathbf{y} - \hat{\mathbf{y}})$ be the models constructed accuracy criterion, Q be the models accuracy criterion values set for proper $\hat{\mathbf{y}}$ matrix estimation, $Q \in \Re^1$, $|Q| = n$, ω be the records set (patients set) where $Q(\mathbf{y} - \hat{\mathbf{y}})$ criterion values are to be found within the bounds of the given error interval (*-0.051, 0.051*), $\omega \subseteq X$.

Let exist mapping F: $Z \rightarrow Y$. The goal is to find such a mapping F^* which guarantee finding the distribution for the accuracy criterion $Q(\mathbf{y} - \hat{\mathbf{y}})$ values $\varepsilon_j \in Q$, $j=1,2,\ldots n$, that as many as possible theirs number are to be found within the bounds of the given error interval (*-0.051, 0.051*) with 0,05 significance level.

2.2 Problem solution

This paper represents generalization of the problem statement in [1]. There were two studies directions in [1]:

1. Construction of the forecasting models pair like:
   - Models without trend:

$$\hat{\mathbf{y}}_1 = f_1(\boldsymbol{\theta}_1, \mathbf{z}_1, \mathbf{z}_2, \mathbf{z}_3) \quad (1)$$

$$\hat{\mathbf{y}}_2 = f_2(\boldsymbol{\theta}_2, \mathbf{z}_1, \mathbf{z}_2, \mathbf{z}_3) \quad (2)$$

   - Models pair $\hat{\mathbf{y}}_1$, $\hat{\mathbf{y}}_2$ as the sum of the trend model (3) and model of the remains (4) и (5).

   Trend model looks like:

$$v(k) = \boldsymbol{\theta}^T \psi(k) \quad (3)$$

   where $\psi(k)$ is the functional transformations vector (k is the discrete time moments), whose elements are:

$$\psi_i(k) = \left(\frac{1}{k}, \sqrt[3]{k}, \frac{}{\sqrt[3]{k}}, 1, k, k^2, k^3, \ldots \right)$$

   Model of remains looks like:

$$\hat{\mathbf{v}}_1 = f_3(\boldsymbol{\theta}_3, \mathbf{u}_1, \mathbf{u}_2, \mathbf{u}_3) \quad (4)$$

$$\hat{\mathbf{v}}_2 = f_4(\boldsymbol{\theta}_4, \mathbf{u}_1, \mathbf{u}_2, \mathbf{u}_3) \quad (5)$$

   where remains were calculated as:

$$\mathbf{u}_k = \mathbf{z}_k - v(k), \; k = 1, \ldots, 3$$

$$\mathbf{v}_1 = \mathbf{y}_1 - v(k), \; k = 4$$

$$\mathbf{v}_2 = \mathbf{y}_2 - v(k), \; k = 5$$

2. Construction of the second forecasting model
   - without adaptive forecast as (2) (or (5) in the case with the use of the trend model)
   - with adaptive forecast as (6) (or (7) in the case with the use of the trend model):

$$\hat{\mathbf{y}}_2 = f_2(\boldsymbol{\theta}_2, \mathbf{z}_1, \mathbf{z}_2, \mathbf{z}_3, \hat{\mathbf{y}}_1) \quad (6)$$

$$\hat{\mathbf{v}}_2 = f_4(\boldsymbol{\theta}_4, \mathbf{u}_1, \mathbf{u}_2, \mathbf{u}_3, \hat{\mathbf{v}}_1) \quad (7)$$

where $\boldsymbol{\theta}_i$ is the parameters vector, which enters in the models linearly, $i = 1,\ldots,4$.

The paper continues investigation in order to get more accurate forecasting results for examination set. It is have also checked of hypothesis about preference of the selection the best models pair $\hat{\mathbf{y}}^* = (\hat{\mathbf{y}}_1^*, \hat{\mathbf{y}}_2^*)$ which minimize $Q(\mathbf{y}-\hat{\mathbf{y}})$ criterion using W set (b) and c) schemes – modified TSMDA) over its selection minimizing $Q(\mathbf{y}-\hat{\mathbf{y}})$ using C set (a) scheme – basic TSMDA) with the purpose of maximizing the power of the ω set, $\omega \subseteq X$, $|\omega| = n_\omega$.

There was utilized TSDMA for the $\hat{\mathbf{y}}_1^*, \hat{\mathbf{y}}_2^*$ forecasting models construction in [1], whose basic idea consist in implementation the next (basic) scheme of W ($W_{1,2}$) active set division in two stages. Set subscript indicates the models number(s) this set should be used for constructing of.

a) The basic TSDMA scheme described in [5] consist in such two stages realization: at the first stage $W_{1,2}$ set is divided into $U_{1,2}$ training and $C_{1,2}$ examination subsets, $U_{1,2}$ subset is used for models construction using GMDH, $C_{1,2}$ – for models evaluation and selection of theirs best pair by the $Q(\mathbf{y}-\hat{\mathbf{y}})$ criterion (diagnostic criterion in our task [1]). At the second stage $U_{1,2}$ subset is divided into A learning and B testing subsets twice: one for each forecasting model. Subsets A_1, A_2 are used for competitive models structures parameters estimation while constructing corresponding forecasting models. Subsets B_1, B_2 – for corresponding best structures selection using external GMDH criterion (see fig.1).

The paper proposes two new TSDMA modifications (schemes) for the forecasting models construction. Let's describe just these schemes differences from basic one.

b) In the basic scheme both forecasting models are built using the same $W_{1,2}$, $U_{1,2}$ data sets. The forecasting models quality evaluation is carried out utilizing the same $C_{1,2}$ subset. Unlike basic scheme this modification do not limit the algorithm freedom in models construction: both models can use any records of the W set in the construction process, i.e. they have individual W_1, W_2 sets and U_1, U_2, C_1, C_2 subsets. Subsets U_1, U_2, C_1, C_2 may be differ either in records number or in records composition. In this scheme the forecasting models quality evaluation is carried out using the W set by the diagnostic criterion (see fig. 2).

c) This scheme differs from the basic one just by the models pair quality evaluation that is carried out using whole W set (see fig. 3).

The division sets study in utilization of complete or incomplete information (incomplete arguments composition) in $W_{1,2}$, W_1, W_2, and $U_{1,2}$, U_1, U_2 sets is also expediently. This study represents particular interest as long as it is supposed to reduce the patients' examination time and TDR measurements number, and is important how much may become worse models accuracy. The *Table 1* shows complete and incomplete information for all sets and subsets mentioned above.

Table 1. Arguments compositions for complete and incomplete information

	$W_{1,2}$	W_1, U_1	W_2, U_2	
■ Complete information	■ $\mathbf{z}_1, \mathbf{z}_2, \mathbf{z}_3, \mathbf{y}_1, \mathbf{y}_2$	■ $\mathbf{z}_1, \mathbf{z}_2, \mathbf{z}_3, \mathbf{y}_1$	■ Adapt. forecast ■ Nonadapt. forecast	■■ $\mathbf{z}_1, \mathbf{z}_2, \mathbf{z}_3, \hat{\mathbf{y}}_1, \mathbf{y}_2$ ■■ $\mathbf{z}_1, \mathbf{z}_2, \mathbf{z}_3, \mathbf{y}_2$
■ Incomplete information	■ $\mathbf{z}_1, \mathbf{z}_2, \mathbf{z}_3$	■ $\mathbf{z}_1, \mathbf{z}_2, \mathbf{z}_3$	■ Adapt. forecast ■ Nonadapt. forecast	■■ $\mathbf{z}_1, \mathbf{z}_2, \mathbf{z}_3, \hat{\mathbf{y}}_1$ ■■ $\mathbf{z}_1, \mathbf{z}_2, \mathbf{z}_3$

Besides quasi-optimal (ρ^2-proportional) this paper investigates such division:

- ■ Similar "by dispersion" (SD) division;
- ■ Dissimilar "by dispersion" (DSD) division.

Figures 1,2,3 describe different divisions application in TSDMA schemes.

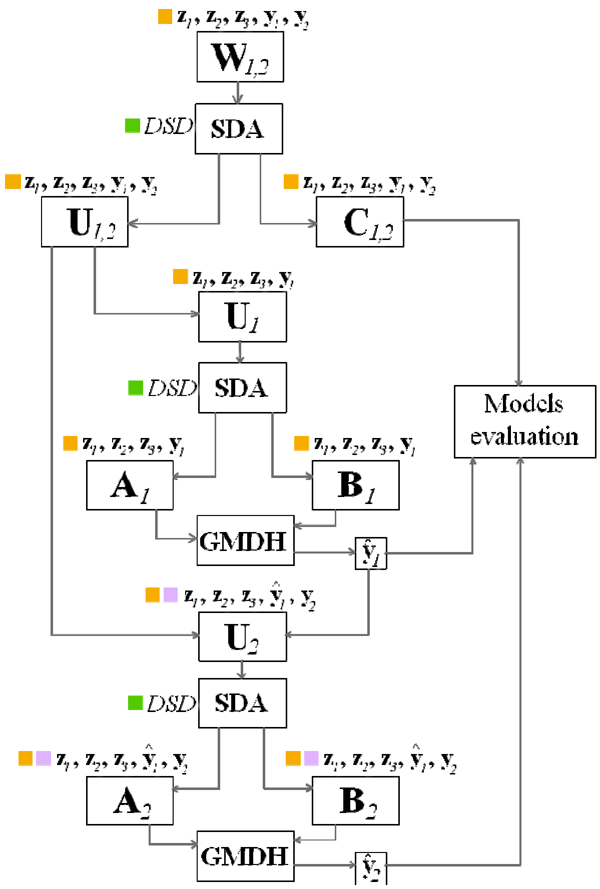

Figure 1. a) Basic TSDMA scheme. This variant uses complete information, dissimilar “by dispersion” division and adaptive forecasting scheme

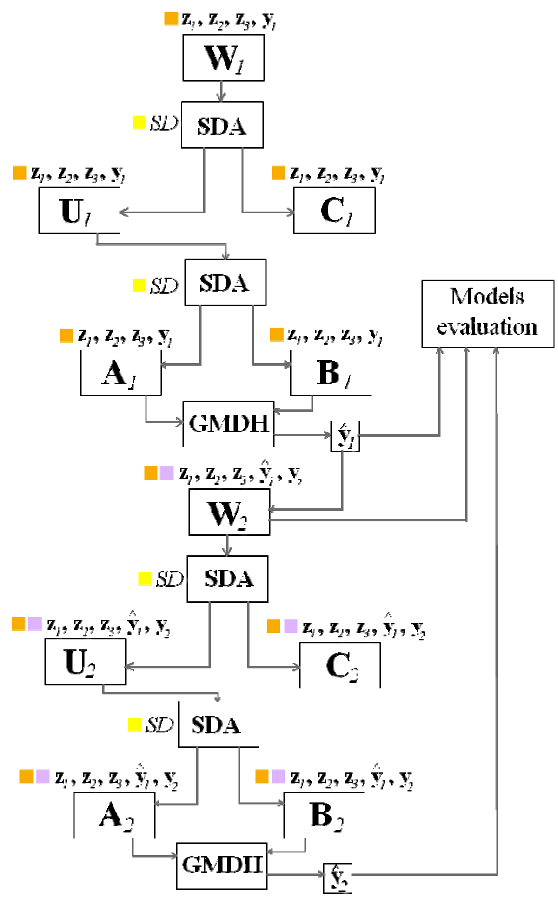

Figure 2. b) modified TSDMA scheme. This variant uses complete information, similar “by dispersion” division and adaptive forecasting scheme

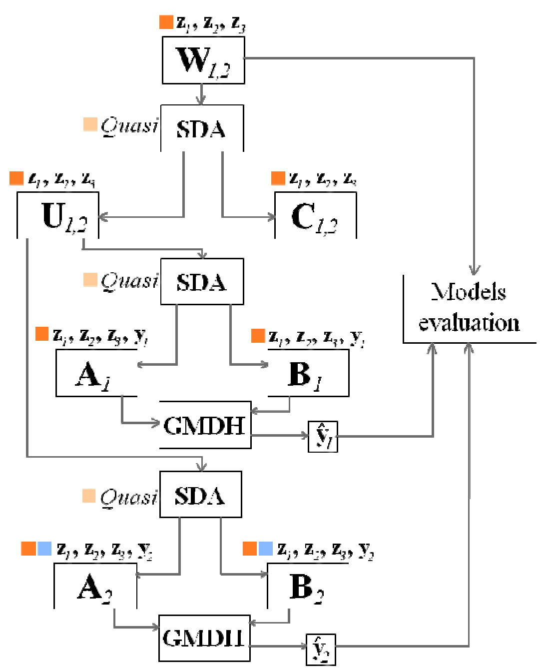

Figure 3. в) modified TSDMA scheme. This variant uses incomplete information, quasi-optimal division and nonadaptive forecasting scheme

Thereby here are such a variations of input parameters for each a), b), c) TSDMA scheme:

1) 2 variants of utilizing initial data in the sets dividing process:
- complete information;
- incomplete information.

2) 3 variants of the set division algorithms:
- Dissimilar "by dispersion" (DSD) division;
- Similar "by dispersion" (SD) division;
- ρ^2-proportional division (Quasi);

3) 2 variants of the schemes for the second model construction:
- with adaptive forecast;
- without adaptive forecast.

4) 2 ways of model construction:
- with trend;
- without trend.

In the lump here are 2·3·2·2 = 24 models pair variants for each scheme.

3 Results

We've used modified version of Multistage Algorithm with Combinatorial-Selection Orthogonalized factors (MACSO) as GMDH algorithm in each mentioned scheme. Such a MACSO input parameters were vary while each models pair (total pairs number is 24·3 = 72) is constructed:

1) the stages number (the iterations number);
2) the adjustments number (number of multiplicative terms);
3) the solution freedom of choice;

Models pairs which have the equal records number n_ω (the records number which models pair satisfy given ±5.1% error interval) were appeared. Selection of the best pair is carried out by such characteristics:

1) n_{out} is the records number of W set where models pair have unsatisfactory error level (>20%);
2) $Q(\hat{z}_3, \hat{z}_4)_{max}$ (%) is the maximal diagnostic criterion deviation, calculated for W set while the forecasting models evaluation is carried out;
3) models complexity.

Actually whole W set is used in the models construction process as the best models pair final selection is realized by sequential use of the diagnostic criterion utilizing C or W data set. Thus we should say about data approximation of W set by obtained models pair.

3.1 Approximation results

The count of the patients number is carried out via $q \in \{0,1\}$ indicator function. These models describe initial data satisfying given error level. The $\hat{\mathbf{y}}^*$ best models pair selection in corresponding TSDMA schemes is realized by such a criteria:

In a) scheme criterion looks like:

$$i_C^* = \arg\max_{i \in C,\, \ell \in \Omega} \left(\frac{1}{n_C} \sum_{i \in C} q_{i\ell} \right), \quad q_{i\ell} = \begin{cases} 1, & \varepsilon_i(p_\ell) \in (a,b) \\ 0, & \varepsilon_i(p_\ell) \notin (a,b) \end{cases},$$

and in b) and c) schemes like:

$$i_W^* = \arg\max_{\ell \in \Omega} \left(\frac{1}{n_W} \sum_{i \in W} q_{i\ell} \right), \; q_{i\ell} = \begin{cases} 1, & \varepsilon_i(p_\ell) \in (a,b) \\ 0, & \varepsilon_i(p_\ell) \notin (a,b) \end{cases}.$$

where p_ℓ is the $W = A \cup B \cup C$ set division parameter for the proper $\ell \in \Omega$, $|\Omega| = L$, Ω is the all possible division variants set when W set two-stage division process takes place.

Numerical experiments showed that variant with the use of the best models pair selection utilizing whole W set in comparison with using its part – C set, leads to getting more accurate models, since $i_C^* < i_W^*$.

More preferable variant selection between competitive ones in proposed studies directions is realized by means of two averaged criteria:

- $Aver_{n_\omega / n_W}$ – averaged percentage $n_\omega / n_W \cdot 100\%$ by the models pair set which use corresponding variant of study direction;
- $Aver_{n_{out} / n_W}$ – averaged percentage $n_{out} / n_W \cdot 100\%$ by the models pair set which use corresponding variant of study direction.

Table 2 shows comparative analysis results of mentioned schemes. As one can see, b) scheme is more accurate then c); a) scheme turns out to be the most inaccurate.

Table 2. Comparison of a), b) and c) schemes while data approximation

Scheme	$Aver_{n_\omega / n_W}$	$Aver_{n_{out} / n_W}$
a)	51 %	10,6 %
b)	65 %	5,56 %
c)	62 %	6,94 %

Table 3 shows three schemes analysis results for all mentioned studies directions. Mark $\succ$ means that variant staying on the left side is to be found more accurate then another one by both criteria. Empty cell means variants incomparability due to different signs between them in mentioned criteria.

Table 3. Approximation results of three TSDMA schemes for each studies direction

Scheme	Information	Division algorithm
	Complete, Incomplete	DSD, SD, Quasi
a)		SD $\succ$ Quasi, SD $\succ$ DSD
b)	Incomplete $\succ$ Complete	SD $\succ$ DSD, Quasi $\succ$ DSD
c)		SD $\succ$ DSD, Quasi $\succ$ DSD
	Model construction scheme for second forecasting model	Ways of models pair construction
	Adaptive, Nonadaptive	With trend, Without trend
a)	Adaptive $\succ$ Nonadaptive	With trend $\succ$ Without trend
b)		With trend $\succ$ Without trend
c)	Adaptive $\succ$ Nonadaptive	With trend $\succ$ Without trend

At the same time the best approximating models were obtained for corresponding schemes with such variants in studies directions:

a) scheme: Complete information, DSD, Adaptive scheme, Models with trend
$n_\omega / n_C \cdot 100\% = 7/8 \cdot 100\% = 87{,}5\%$;
$n_{out} = 1$, $Aver_{n_{out} / n_C} = 1/8 \cdot 100\% = 12{,}5\%$.

b) scheme: Incomplete information, Quasi, Adaptive scheme, Models with trend;
$n_\omega / n_W \cdot 100\% = 14/18 \cdot 100\% = 77{,}8\%$;
$n_{out} = 0$. $Aver_{n_{out} / n_W} = 0\%$.

c) scheme: Complete information, Quasi, Adaptive scheme, Models with trend;
$n_\omega / n_W \cdot 100\% = 14/18 \cdot 100\% = 77{,}8\%$;
$n_{out} = 3$, $Aver_{n_{out} / n_W} = 3/18 \cdot 100\% = 16{,}7\%$.

For the models verification purpose were used randomly selected checking *D* set, with the power in 10 records.

3.2 Verification results using checking set

Table 4 shows all selected studies directions analysis results for each scheme. Meanwhile preference is given to selection the best variant using the same criteria $Aver_{n_\omega / n_D}$ и $Aver_{n_{out} / n_D}$.

Table 4. Verification analysis results of three TSDMA schemes for all studies directions

Scheme	Information	Division algorithm
	Complete, Incomplete	DSD, SD, Quasi
a)	Incomplete $\succ$ Complete	Quasi $\succ$ SD $\succ$ DSD
b)		Quasi $\succ$ SD, Quasi $\succ$ DSD
c)	Complete $\succ$ Incomplete	Quasi $\succ$ SD $\succ$ DSD
	Model construction scheme for second forecasting model	Ways of models pair construction
	Adaptive, Nonadaptive	With trend, Without trend
a)	Adaptive $\succ$ Nonadaptive	With trend $\succ$ Without trend
b)	Nonadaptive $\succ$ Adaptive	Without trend $\succ$ With trend
c)		With trend $\succ$ Without trend

At the same time the best models were obtained for corresponding schemes with such variants in studies directions:

a) scheme: Incomplete information, Quasi, Adaptive scheme, Models with trend;

$n_\omega/n_D \cdot 100\% = 3/10\cdot 100\% = 30\%$;

$n_{out} = 1$, $Aver_{n_{out}/n_D} = 1/10\cdot 100\% = 10\%$.

b) scheme: Incomplete information, Quasi, Nonadaptive scheme, Models without;

$n_\omega/n_D \cdot 100\% = 3/10\cdot 100\% = 30\%$;

$n_{out} = 1$, $Aver_{n_{out}/n_D} = 1/10\cdot 100\% = 10\%$.

c) scheme: Incomplete information, Quasi, Adaptive scheme, Models with trend;

$n_\omega/n_D \cdot 100\% = 6/10\cdot 100\% = 60\%$;

$n_{out} = 2$, $Aver_{n_{out}/n_D} = 2/10\cdot 100\% = 20\%$.

Table 5 shows a), b), c) schemes comparative analysis results using checking *D* set. As one can see, c) scheme is the most accurate, then goes b) scheme, and then a) one.

Table 5. Comparison of the a), b) and c) schemes while model verification

Scheme	$Aver_{n_\omega/n_W}$	$Aver_{n_{out}/n_W}$
a)	20,8 %	30 %
b)	20,8 %	29,6 %
c)	23,8 %	30 %

Table 6 shows generalized analysis results all over the schemes and studies directions.

Table 6. Verification analysis results all over the studies directions

Information	Division algorithm
Complete, Incomplete	DSD, SD, Quasi
	Quasi $\succ$ SD $\succ$ DSD
Model construction scheme for second forecasting model	Ways of models pair construction
Adaptive, Nonadaptive	With trend, Without trend
	With trend $\succ$ Without trend

As one can see, generalized analysis results in *Table 6* are conformed with variants corresponding to the best models pair – the pair in c) scheme mentioned a little bit earlier. Results in *Table 5* unambiguously indicate that the best models pair answers c) scheme, that actually corresponds partial selection result of the best models pair.

Let's carry out statistical analysis of the best models pair selected. Suppose that statistical properties of the models error values $\varepsilon_W \in E_W$ and $\varepsilon_X \in E_X$ of randomly selected *W* and *X* sets from general totality are the same, i.e. distribution functions, expectation values and dispersion values σ^2 are equal. Such hypothesis is checked . If empirical density function curve will be close to symmetrical (or Gaussian) distribution law by mathematical statistics criteria, then the conclusions using 3σ rule obtained utilizing *W* set are supposed to be true for the *X* set. Notably if some statement for the linear by parameters $\boldsymbol{\theta}$ model $\hat{\mathbf{y}}^* = f(\boldsymbol{\theta},\mathbf{z})$ obtained using *W* set is true then it would also be true for the *X* set. Thereby this model deviation from a true TDR values for the main group of the *X* set patients would also be found in the error interval (a, b) with 0,05 significance level.

The hypothesis checking results are presented in histograms (see fig. 4, fig. 5). There are built the frequency histogram of the $P_W(\varepsilon_n \in (a_i, b_i)), n=1,2,\ldots,n_W$ and $P_X(\varepsilon_n \in (a_i, b_i)), n=1,2,\ldots,n_X$ for the best models pair (Incomplete information, Quasi-optimal division, Adaptive scheme, Model with trend). As one can see, frequency functions (discrete type of destiny function) for both *W* and *X* sets are close to normal distribution law. The standard deviations are $\sigma_{\varepsilon_W} = 5.32$ and $\sigma_{\varepsilon_X} = 9.52$. Although $\sigma_{\varepsilon_W} < \sigma_{\varepsilon_X}$ but using the 3σ criterion ($3\sigma_{\varepsilon_W} = 15{,}95$, $3\sigma_{\varepsilon_X} = 28{,}57$) most of the patients belonging to the ω set are to be found in the given error interval (*-0.051, 0.051*) with significance level as low as 0,04. So we have such results: 100% patients of *W* set and 96% patients of *X* set (just one record from the 28 did hit the interval) are to be found within the 3σth interval.

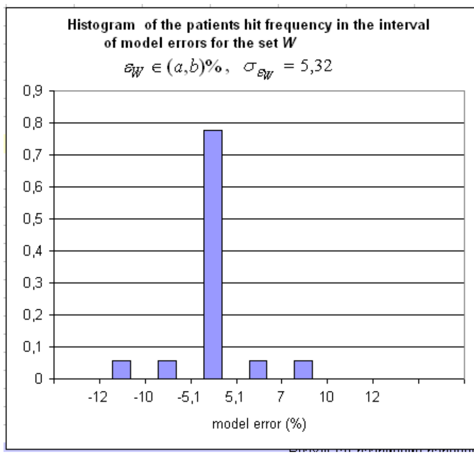

Figure. 4

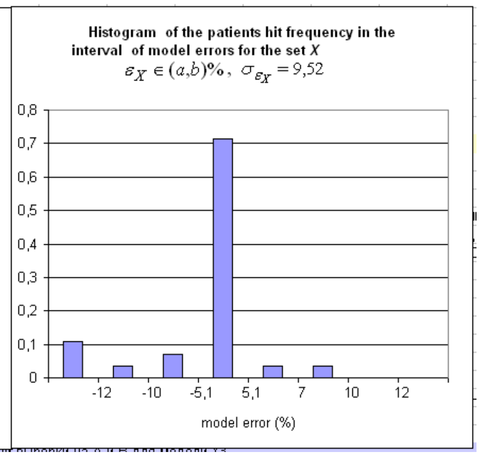

Figure. 5

4 Conclusion

We came to such conclusions while resolving this problem:

- When constructing approximating models it is preferable to use :
  - Quasi-optimal set division algorithm;
  - Adaptive scheme for construction of the second forecasting model;
  - Construction of the models pair utilizing trend model;

  b) scheme – scheme with models construction and evaluation utilizing whole *W* set.

- When constructing forecasting models it is preferable to use:
  - Quasi-optimal set division algorithm;
  - Construction of the models pair utilizing trend model;
  - Incomplete set of arguments when set dividing process takes place;

  c) scheme – scheme with models construction utilizing *U* set, and evaluation utilizing *W* set.

- The variant with the best models pair selection by diagnostic criterion using whole *W* set leads to getting more accurate results either for construction approximating and forecasting models.

- The best models pair statistical analysis results denote about its good forecasting properties. But it is necessary to have a confirmation of different models forecast quality of the best models pairs set for the mass data.

References

[1] Nina Kondrashova, Andriy Pavlov.: GMDH-based Forecasting of the Test Results of Blood Samples in Task of the Effective Medicines Estimation. *Proceedings of 2nd international conference on Inductive modeling, Sept.15-19, Kyiv, p. 295-300, 2008,* ISBN 978-966-02-4889.

[2] Kondrashova N.V., Pavlov A. V.: Forecasting models for the estimation of thiol-disulfide ratio by the three-stage sample division. *Control Systems and Computers*, 2008, №3. (in Russian).

[3] Stepashko V.S. The structural identification of forecasting models in conditions of the design of experiments.– *Automatica, 1992, №1, p. 26-35.* (in Russian).

[4] Kondrashove N.V. Forecasting models accuracy comparison when vary dividing set. Coll. Modeling, *2008. – C.128-136.* (in Russian)

[5] Nina Kondrashova, Andriy Pavlov, Yaroslav Pavlov.: Optimization of Forecasting Models for Testing Blood Samples by Estimation of Thiol-Disulfide Diagrams. *IWIM 2007 Workshop proceedings, 160-165, 2007*