Optimization of Forecasting Models for Testing Blood Samples by Estimation of Tiol-Disulfid Diagrams

Nina Kondrashova[1], Andriy Pavlov[2], Yaroslav Pavlov[2]

[1]*International Research and Training Center of Information Technologies and Systems of the National Academy of Sciences of Ukraine, Ukraine*

[2]*Kiev polytechnic institute, National technical university of Ukraine , Ukraine*

nkondrashova@ukr.net, me_ovechka@bigmir.net

Abstract. *Constructing the forecast models for values of tiol-disulfide ratios in blood samples in several measurement points is considered for decreasing the time of patients' examination. Models are obtained by the GMDH algorithms. The number of examination measurement points in which models have feasible error is maximized by optimization of initial data sample division, by adaptive forecast, and sequential use of selection criteria. The criteria and results of numeric experiments are given. Application of difference models with an adaptive forecast and algorithm of modeling with the two-stage division of initial sample are effective for the variables forecast.*

Keywords

Forecasting model, GMDH algorithm, sample division,
adaptive forecast, tiol-disulfide ratio

1 Introduction

The reliable model definition is one of the nowadays issue which arise at medical diagnostics. Costliness of preparations, medical service, considerable level of poverty of population and many others stimulates works for reduction the time of a patients' examination. The results of researches of blood samples of oncological patients were the initial data for modeling in our work. The accumulated statistical testing of blood samples database is used for building models. It consists of approximately 300 patients tested by medicinal preparations (about 130 preparations with different dosage) for a few years.

Now tiol-disulfid ratio (TDR) is measuring in five points of time interval (see figure 1) but it is assumed to reduce the amount of measuring points down to three ones (see figure 2) in future. Figure 1 represents the initial data of the work sample of rationed TDR values obtained in five points of the time interval $j = \{0,..,4\}$ for the blood samples of $N_U = 8$ patients. Figure 2 shows this information for three measuring points ($j = \{0,1,2\}$ these points are connected by lines) and the values in 4*th* and 5*th* points (showed as the isolated points) which are obtained by using forecast models for this patients' group.

Possibility of receiving a model for *any* point of measuring was considered in [1]. We could define which one of the measuring points can be substitute by the model value obtained using the proper model based on the value of a diagnostic criterion for which the possible error value ε is given

by an expert. And then the proper models are used for calculating values of the diagnostic criterion for next patients in future.

2 Theoretical Part

In this work the optimal forecasting models for two last points of TDR measuring are describing the following expressions:

$$\hat{z}_k = f(x_0,...,z_{k-j}), \quad k=\overline{3,4}; \quad J=\overline{2,3}; \quad (k-j)=\overline{1,J} \tag{1}$$

Here $\hat{z}_k$ is calculated using real measuring in previous 3 points. Thus it is possible more faster make solutions about treatment because of the observation time interval is reduce down to 1 hour.

- $z_{k-j}=\dfrac{y_{k-j}}{x_{k-j}}$ are relative TDR values in the points of time interval from 0 hour up to 1 hour, $j=\overline{(k-1),(k-2)}$;
- x_{k-j} are TDR values of control curve (reaction of blood sample of appropriate patient on the distilled water);
- y_{k-j} are TDR values of the blood sample cultivated by the certain dosage of investigated medicinal preparations;
- $k=\overline{3,4}$ correspond to the time counts 3 hours and 24 hours.

The terms of patient examination decrease from 24 hours down to 1 hour by using such models.

For the fourth point of measuring a model is defined by using an adaptive forecast [2] at $j=\overline{(k-1),(k-3)}$, $z_3 \overset{\Delta}{=} \hat{z}_3$ in a formula (1).

Algorithms of GMDH, optimization of initial sample division $W = A \cup B \cup D$ to *A* teaching, *B* checking and *D* examination subsamples and also *successive application of selection criteria* is used for the construction of models (1). Sample W contains the nonnormable TDR values x_0 in zero point and the normable TDR values z_j in points $j=\overline{1,4}$ obtained by using blood samples with certain medicinal preparation for the patients' group.

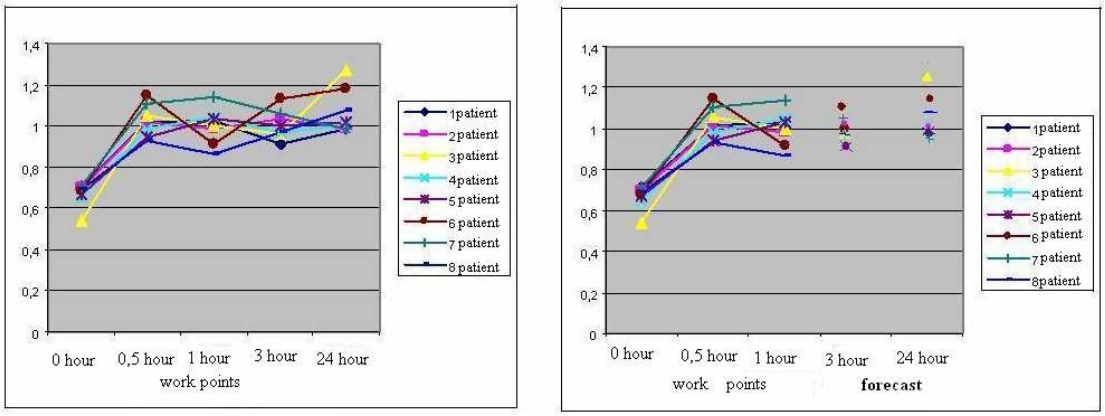

Fig. 1. Initial sample of the normable TDR values.

Fig. 2 Initial sample of the normable TDR values for first three time points and TDR values for two last time points which calculated by models

The modeling was carrying out based on application of ASTRID package algorithms (COMBI, GMDH, MULTI) [3] each of them has own features and limitations as applied to this task. We came to the necessity of expedient determination of length of $U = A \cup B$ working sample while modeling. Accuracy of the obtained models depends on this parameter and others «not factors». The common limitation of using all indicated algorithms is absence of optimization of initial sample division to

subsamples *A*, *B* and *D*. So there has arisen the problem related to the selection of division method, automation and optimization of samples and subsamples division.

The ρ-proportional (*«*quasi-optimal») division which minimizes the norm of information matrices difference is used for determination of examination and checking subsamples and looks as:

$$p^*_{\min}(N_v) = \arg \min_{\rho^2_\ell \neq 0, \ell=\overline{1,L_v}} \left\| \rho^2_\ell \chi_{1v_\ell} - \chi_{2v_\ell} \right\|, \quad v=\overline{1,2},\ \chi_{iv_\ell} = \mathbf{X}^T_{iv_\ell} \mathbf{X}_{iv_\ell},\ i=\overline{1,2}.$$

Matrices dimensions are $\dim \chi_{iv_\ell} = (J+2)\times(J+2)$, $\dim \mathbf{X}_{iv_\ell} = n_i \times (J+2)$ for *i*th subsamples $i=\overline{1,2}$. Set of divisions variants equals the number of simple combinations:

$$L_v = \sum_{n=n_{\min_v}}^{n_{\max_v}} C^n_{N_v} \tag{2}$$

The value v determines subsample (sample) number which is divided in two subsamples. For the proper matrices we will write down:

$\dim \mathbf{X}_{1v_\ell} = n\times(J+2),\ \dim \mathbf{X}_{2v_\ell} = (N_v - n)\times(J+2)$.

2.1 Problem statement

The set of models is a linear in $\theta_0,...,\theta_m$ parameters difference models in the examined algorithms. For description of class of models we will write down expression:

$$\hat{z}_k = a_0 + b_0 x_0 + \sum_{j=k-1}^{k-2} a_{k-j} z_{k-j} + x_0 \sum_{j=k-1}^{k-2} b_j z_{k-j} + \sum_{j=k-1}^{k-2} \sum_{i=k-1}^{k-2} a_{ij} z_{k-j} z_{k-i} + ..., \quad j \neq i, \quad k=\overline{3,4}$$

In the view of expert a model is good if it satisfies to the given accuracy of diagnostic criterion calculation:

$$H_n = \sum_{k=\overline{1,4}} z_{k,n} - 3$$

Therefore estimated diagnostic criterion of n *th* patient ($n \in \Omega_D$) depends on total accuracy of measuring models:

$$\hat{H}_n = \sum_{k=\overline{1,2}} z_{k,n} + \sum_{k=\overline{3,4}} \hat{z}_{k,n} - 3$$

It must be within the limits of possible errors on the possibly greater number of points of examination subsample where Ω_D is set of points of examination subsample which depends on the type of division (here the *D* index means belonging to the examination set with the n_D number of elements). Error of diagnostic criterion determination for a group of n_D patients is:

$$\Delta_D = 100\% \sum_{n=\overline{1,n_D}} \left| H_n - \hat{H}_n \right| \Big/ \sum_{n=\overline{1,n_D}} H_n$$

It mustn't exceed ε=5%. It is necessary to take into account when selecting every model using examination points $n \in \Omega_D$:

$$Cr_D(n_D, \hat{\mathbf{z}}) = 100\% \sum_{n=\overline{1,n_D}} \sum_{k=\overline{3,4}} \left| z_{k,n} - \hat{z}_{k,n} \right| \Big/ \sum_{n=\overline{1,n_D}} \sum_{k=\overline{3,4}} z_{k,n},\ \hat{\mathbf{z}} = (\hat{z}_3, \hat{z}_4)$$

It is solving the following task of optimal eight determination $\left\{p^*_1, N^*_U, p^*_2, n^*_B, s^*_3, \boldsymbol{\theta}^*_3, s^*_4, \boldsymbol{\theta}^*_4\right\}$ by maximization n_d number of examination points (elements of set Ω_d, $\Omega_d = \Omega_N \setminus \Omega_{N_U}$) in which models have feasible error. Here s^* is the optimal model complexity, $\hat{\boldsymbol{\theta}}_{s^*}$ is the optimal model parameters vector ($\dim \hat{\boldsymbol{\theta}}_{s^*} = 1\times m_{s^*}$), N^*_U is the working sample length:

$$\left\{p^*_1, N^*_U, p^*_2, n^*_B, s^*_3, \boldsymbol{\theta}^*_3, s^*_4, \boldsymbol{\theta}^*_4\right\} = \arg \max_{\Omega_{N_U} = \Omega_N \setminus \Omega_D, \Omega_d \subseteq \Omega_D} n_d, \quad n_d = |\Omega_d|,\ N_U = |\Omega_{N_U}|$$

$$n_d = \arg \min_{n_i \in \Omega_D} \left| Cr_D(n_i, \hat{\mathbf{z}}^*) - \varepsilon \right|, \ \hat{\mathbf{z}}^* = \left(\hat{z}_3^*, \hat{z}_4^* \right) \tag{2}$$

2.2 Two-stage division modeling algorithm

Model with optimum complication s is checked up using information of examination subsample obtained by GMDH *algorithms* after division of *W* initial sample to *U* working subsample and *D* examination subsample (v =1) and then division of obtained working subsample to *A* learning subsample and *B* checking subsample (v =2).

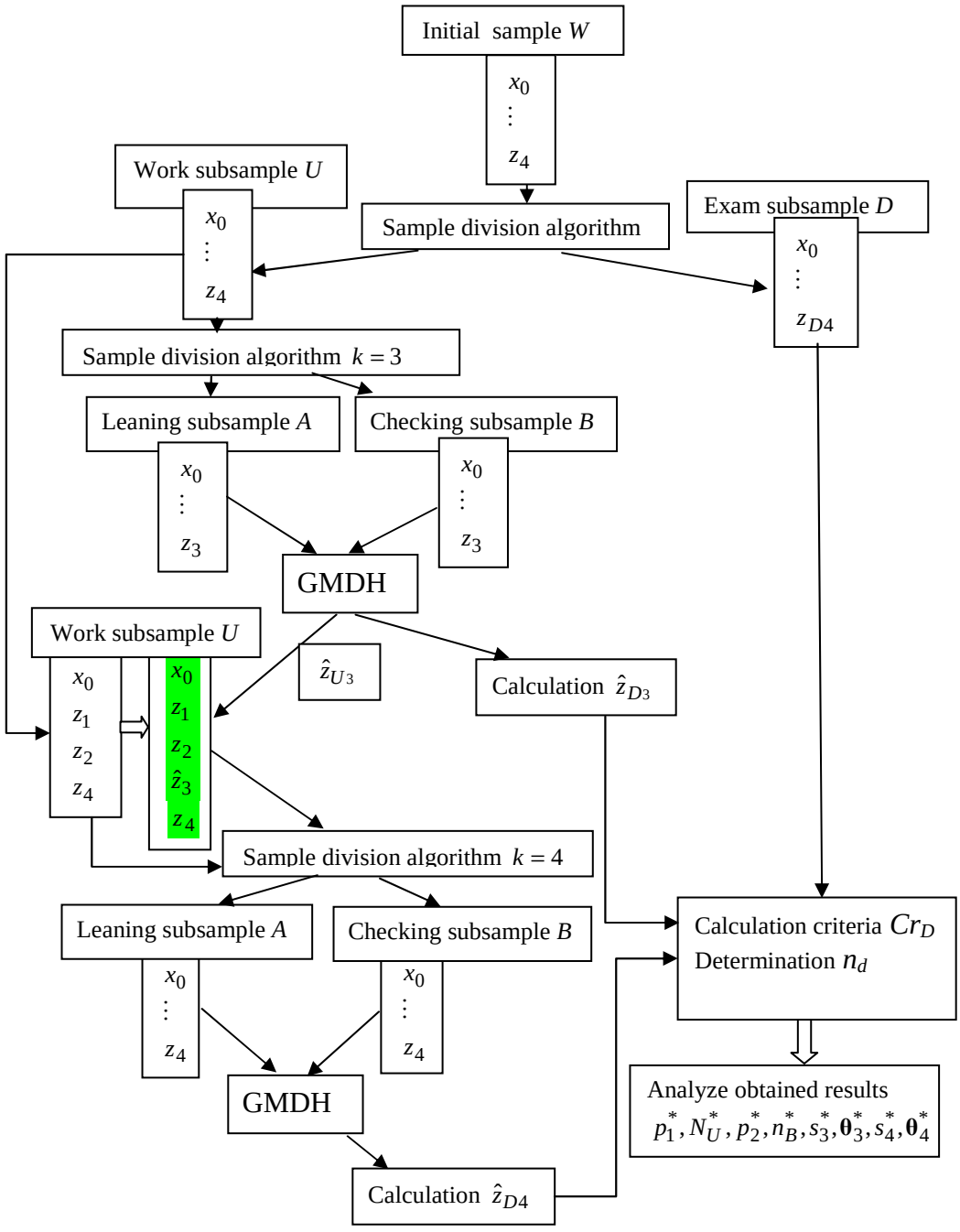

Fig. 3 Block-chart of two-stage division modeling algorithm.

Information of *W* initial sample ($N_1 \overset{\Delta}{=} N$, $n = N_U$ in a formula (2)) is used in order to get *U* working and *D* examination subsamples (here index $v = 1$) for calculation of the χ_{11} and χ_{21} informative matrices.

Information of obtained *U* subsample ($N_2 \overset{\Delta}{=} N_U$, $n \overset{\Delta}{=} n_B$ in a formula (2)) is used for obtaining of *A* and *B* subsamples (here index $v = 2$) for calculation of the χ_{12} and χ_{22} informative matrices.

The procedure repeats oneself for the all sets of divisions. The optimal p_1 set of lines, their N_U amount and also $\overline{p}_1$, $n_D = N - N_U$ are the results of the first stage.

The result of the second stage are p_2, n_B and $\overline{p}_2$, $n_A = N_U - n_B$.

The $\overline{p}_v^*$, p_v^*, $v = \overline{1,2}$ rows sets of optimal division, $N_U^*, n_D^*, n_A^*, n_B^*$ optimal lengths of subsamples, and s_k^* optimum complexity of $\left\{\hat{z}_k^*\right\}$ models (i.e. structures with the $m_{s_k^*}$ arguments, and $\boldsymbol{\theta}_k^*$ parameters where $k = \overline{3,4}$) are analyzed using the results of model verification of an examination subsample. The maximum of n_d number of examination points is analyzed for the best models by $Cr_D(n_d)$ criteria satisfying given error.

The block-chart represents the research possibility of models without an adaptive forecast and with an adaptive forecast (in a figure 3 the proper rectangle is painted) for the different variants of working sample division. The dimensions of informative matrices are $\dim \chi_{i2} = 4 \times 4$ in the first variant and $\dim \chi_{i2} = 5 \times 5$ in the second variant

2.3 Results

Our analysis is based on research of random sample which consists of 19*th* patients' TDR values obtained of influence on the blood samples by immunofan preparation with the dosage 1 ml. As a result were obtained the models $\hat{z}_{k,n}$ $k = \overline{3,4}$ using GMDH (COMBI) algorithms in the class of linear structures with a selection of the regularity criterion using the optimal divided sample ($N_U = 8$ at *N*=19) with a maximal number n_d=5 points. Table 1 shows the advantage of adaptive forecast.

Tab.1. Number of models and their accuracy using the quasi-optimal division

	without an adaptive forecast	adaptive forecast
$Cr_D(n_d)$	4,28	4,82
n_d	1	5

Conclusion

The following conclusions are possible in view of the above results of numerical experiments:

- Volumes and contents of working, teaching and checking subsamples influence on accuracy of the built models.
- Application of difference models with an adaptive forecast and algorithm of modeling with the two-stage division of initial sample is effective to forecast in case of small number of variables and relatively large number of points (*N*>*J*).