Joanna BOBKOWSKA
WEST POMERANIAN UNIVERSITY OF TECHNOLOGY OF SZCZECIN, FACULTY OF COMPUTER SCIENCE AND INFORMATION TECHNOLOGY,
ul. Żołnierska 49, 71-210 Szczecin

Comparison of the results of time series prediction obtained with the classical GMDH algorithm and the modified method containing sensitivity functions

Phd student Joanna BOBKOWSKA

I studied at the Faculty of Mathematics and Computer Science of Adam Mickiewicz University in Poznan. The Master's thesis concerned the mathematical foundations of computing and the Turing machine. Currently I'm dealing with time series analysis, prediction for short sample of input data and simulation of complex systems behavior.

e-mail: jbobkowska@wi.zut.edu.pl

Abstract

The paper presents the results of prediction experiments dealing with the behavior of a complex process containing significant regularity which is modeled by a given time series. In my research I use only a small amount of the input data in order to predict future states of the aforementioned time series using a modified GMDH containing sensitivity functions. It turns out that, for some specific processes, sensitivity functions allow us to obtain more accurate results than the classical GMDH.

Keywords: complex systems, prediction, sensitivity functions, GMDH, partial models, Kolmogorov-Gabor equation, time series, short sample of data.

Porównanie rezultatów predykcji szeregów czasowych uzyskanych za pomocą klasycznego algorytmu GMDH oraz zmodyfikowanej metody GMDH z funkcjami czułości

Streszczenie

Poniższy artykuł przedstawia wyniki eksperymentów dotyczących predykcji zachowania pewnego złożonego procesu zawierającego znaczne regularności, który modelowany jest za pomocą szeregu czasowego. W celu predykcji kolejnych wartości szeregu korzystam jedynie z niewielkiej ilości danych wejściowych stosując zmodyfikowaną metodę GMDH (Group Method of Data Handling) zawierającą funkcje czułości. Metody statystyczne stosowane zwykle w celu ustalenia zależności między poszczególnymi zmiennymi są całkowicie nieprzydatne w warunkach niewielkiej ilości danych wejściowych. Trudno w takich warunkach dostrzec i zbadać regularności szeregu i zależności pomiędzy zmiennymi tego szeregu. Nawet jeśli badany szereg jest szeregiem ze ściśle określoną regularnością, to nie mamy pewności, że ilość próbek, na których ma sposobność pracować badacz jest wystarczająca do określenia wszystkich jego cech. Proces przedstawiony za pomocą pewnego szeregu, może mieć np. składnik cykliczny, który przy małej ilości próbek będzie niewidoczny. Korzystamy więc z narzędzia umożliwiającego uchwycenie wahań analizowanego procesu, jego siły czy kierunku wykorzystywanego między innymi w dyscyplinach zajmujących się sterowaniem procesami. Jednym z takich narzędzi szacujących są właśnie funkcje czułości. Uzyskiwane rezultaty badań pokazują, że zastosowanie funkcji czułości pozwala na otrzymanie dokładniejszych wyników predykcji niż klasyczna metoda GMDH dla pewnych szczególnych zachowań procesu.

Słowa kluczowe: systemy złożone, predykcja, funkcje czułości, GMDH, modele częściowe, równania Kołmogorowa-Gabora, szeregi czasowe, próbki niewielu danych.

1. Introduction

Dealing with the modeling of complex processes which describe the functioning and relationships of certain fragments of reality (be it economics, sociology, biology or physics) and the prediction of their behavior one should initially evaluate the model features and the software required for the simulation. The most popular data mining methods designed for solving some of the above-mentioned issues are based on statistical methods. This means that a researcher with a sufficient amount of experimental data samples (usually a significant quantity of them) first analyses statistical parameters of those processes (the parameter expectations, the statistical dispersions, etc.) [1]. It is believed that the probability theory can work effectively only if the number of time points at which we collect the experimental data is much greater (in practical applications, about 10 times greater) than the number of those at which the created models can be used to control the process. Moreover, among the classical approaches to the analysis of random processes we should mention the theory of the same name – the random process theory. This theory and those mentioned above are also based on the probability theory and various statistical methods [2]. The regression method and the Bayesian networks can be given as an example. Unfortunately, although the regression method allows you to create deterministic models, it does not allow extracting the components which form a complex process.

All of these approaches are effective while the researcher has a sufficient amount of data samples at his disposal. The statistical methods are completely useless if a small amount of input data exists.

In these circumstances, it is difficult to detect and examine regularities in the data series and dependencies between variables of that series. Even if the tested series has a significant regularity, we are not sure that the number of samples with which the researcher has the opportunity to work is sufficient to determine all of its features. For example, the process represented by a certain series may have a cyclic component which could not be found when a small number of samples is tested.

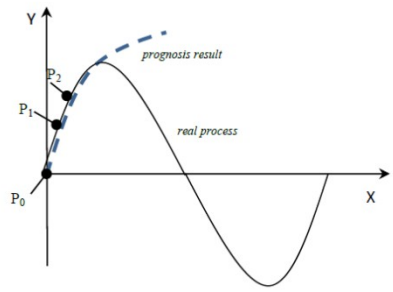

Fig. 1. Chart of a probable prediction run of a cyclic process based upon a small amount of input data
Rys. 1. Wykres prawdopodobnego przebiegu predykcji cyklicznego procesu przy niewielkiej ilości danych wejściowych

Suppose we have a cyclic process described by a function and three samples that accurately describe the value of y variable at three time points P_0, P_1 and P_2 (Fig . 1). It is obvious that it is not possible to determine the periodicity of future development of the process on the basis of only these three experiments. The dotted line illustrates how the process will probably proceed in the future by any prognosis method.

What is needed is a greater number of sample points. However, using a suitable strategy we can greatly reduce the quantity of necessary time points, and obtain satisfactorily accurate results of the prognosis.

Is it possible, being in possession of a small amount of test samples, to build the model, which we will use in the control and prediction of a complex process? The model should use a mathematical tool capable of capturing variations in the process, their strength and direction. Such tools are used, among others, in the disciplines involved in the process control or the design of efficient feedback regulators. Therefore, in order to assess the effectiveness of the controls, we must be able to quantify the relationship between the obtained error, the process and the controller. One such estimating tool is the sensitivity functions [3] which determine the possibility of reducing, minimizing the interferences in a feedback system. The sensitivity functions show us how disturbances affect the operation of a feedback system. So they can be a very promising instrument in forecasting a time series.

2. Modification of the GMDH algorithm

The aim of the research is to create a mathematical tool that allows solving the problem of prediction for a small sample size. One of the algorithms creating a self-organizing model (with extremely high-degree polynomials, gradually complicated polynomial models) which is used in such fields as data mining, knowledge discovery, prediction, complex systems modeling, optimization and pattern recognition, is the GMDH (Group Method of Data Handling) [4,5]. The method was originated in 1968 by Alexey G. Ivakhnenko. It is perfect for complex, unstructured systems when a researcher is interested in obtaining the high input-output relationship only. Ivakhnenko developed the method based on the idea of Rosenblatt's perceptrons (1958) which allowed researchers to build models of a complex system without the assumption of their internal working. The GMDH algorithm creates even the hundredth-degree polynomials where the standard regression sticks in calculations.

To predict the behavior of complex systems simulated by time series, we propose the approach of a modified GMDH algorithm containing the sensitive functions where coefficients of the Kolmogorov-Gabor polynomial (1) are calculated as in the standard GMDH.

The basic GMDH is a procedure of constructing high degree polynomials of the form:

(1)

where m is the number of the base function components. Although it is similar to the high-order regression-type polynomial, the way of constructing these polynomial differs from the standard regression analysis techniques. In fact, it is more similar to the way of natural selection in nature. It is classified as a self-organizing algorithm.

The starting point of our method is Kolmogorov-Gabor polynomial of form (2) which will calculate the future values of the time series.

(2)

To simplify the notation, we will use the characters x, y and z assuming the time series to be three-dimensional. Then by the $X(x, y)$ we understand the value of the x variable in the next time point presented as a partial function of only two variables (x and y) whose previous values are defined. The idea of partial functions where the future values are calculated using a combination of any pair of variables containing the searched one is directly borrowed from the GMDH.

Thus, the future time point of x value can be calculated finding the value of $X(x, y)$ according to the formula (3)

(3)

or (4)

(4)

Similarly, we calculate the forecast of y value as a function of x, y variables by searching for $Y(y, x)$ or function of y and z variables calculating $Y(y, z)$. We use $Z(z, x)$ or $Z(z, y)$ for variable z. So we need to find six (5) models for three variable time series.

(5)

It is necessary to appoint six coefficients for each model in order to find the values of all partial functions. So we solve the system of six equations with six unknown values (6) for $P_0 – P_6$ time points:

(6)

where (for a process ***P***) describes the behavior of the process at the time point .

The contributions of every variable in the values of *x, y, z* in different sections of observation can be significantly different. We can already notice the importance of each variable at the stage of calculating the coefficients (the values of some of them can be close to zero). The sensitivity functions calculated for each partial function contain necessary information which we can use to assess on influence of the process variables at different time points including the analysis of the current and future states of the process to be determined. The classical sensitive functions are used in mathematical modeling to study the efficiency of the models depending on parameters and initial conditions. [6]

Knowing the values of the coefficients we calculate the values of the corresponding sensitivity functions for each partial function. They can be obtained by calculating directly corresponding derivatives of polynomials (3), (4) etc., for every partial function. For partial function $X(x, y)$ we obtain (7).

(7)

Since the sensitivity functions are defined by the partial derivatives of each partial function of local character they have also local character (relate to a range). So the sensitivity (insensitivity – if the values of sensitivity functions are close to zero, the partial function is insensitive to the parameter changes) will depend on the selected time point. The same function may have high values at one range and low at another range.

The GMDH does not allow for splitting the process into separate components but it works on the whole set of variables. The modified method requires the separation of the components of the complex process and evaluation of the contribution of each component to the general process state. In addition, if we want to mathematically determine how to transit one state of the process into another, we can (using the theory of complex processes control terminology [7]) apply the *F* operator in (for simplicity) a three-dimensional metric space (x, y, z) as follows:

. (8)

This operator determines what state the process will have at the next time point P_{i+1}, if the state variables in (x, y, z) space at P_i time point are modified with relative increases , , . We call the problem of determining the actual values of the increases ddx_i, ddy_i, ddz_i which leads the process's variables from (x_i, y_i, z_i) to $(x_{i+1}, y_{i+1}, z_{i+1})$ the **inverse problem**.

In the inverse problem of control of complex processes we need some explicit given values, and we also have to be careful with the representation of uncertain data.

In our experiments we create the function corresponding to the *F* operator (8). The first and second order sensitivity functions (7) are the basis for development of this function. A further strategy for the modified GMDH is as follows:

1. First, we construct the simplified partial functions with first order sensitivity functions. We obtain (9) for the partial function $X(x, y)$. Solving the inverse problem we find the values of the relative increases *ddx* and *ddy* at time points P_0, P_1, P_2, ... calculating the system of linear equations with $X_{i+1}(x, y)$ and $Y_{i+1}(y, x)$ ((9) and (10)).

(9)

(10)

2. Then we create a detailed function using the first and second order sensitivity functions in the form of (11) and (12) and solve the inverse problem again. It means we find the values of the relative increases *ddx* and *ddy* at the time points P_0, P_1, P_2, ...

(11)

+

(12)

3. Finally, we compare the results of the calculations for the simplified and detailed models and choose the model that gives better results for further calculations.

3. The results of the experiments

The experiments with the modified GMDH containing the sensitivity functions were carried out inter alia with three variable time series of form as described in Table 1.

(13)

Tab. 1. The data set with which the experiments were carried out using the GMDH and the modified GMDH, where values of z were obtained using (13)
Tab. 1. Zbiór danych, na którym przeprowadzono eksperymenty z wykorzystaniem zmodyfikowanej metody GMDH, gdzie wartości zmiennej z obliczano ze wzoru (13)

Time points	x	y	z
P_0	1,1	0,69	1,581141
P_1	1,21	0,7935	2,211733
P_2	1,331	0,912525	3,105733
P_3	1,4641	1,049404	4,357086
P_4	1,61051	1,206814	6,061011
P_5	1,771561	1,387836	8,264303
P_6	1,948717	1,596012	10,85823
P_7	2,143589	1,835414	13,40973
P_8	2,357948	2,110726	15,02293
P_9	2,593742	2,427335	14,51877
P_{10}	2,853117	2,791435	11,28116

We calculated the coefficients of the polynomials for each partial function at time points $P_0 – P_6$ solving the system of equations (6). We obtained 6 models (sets of coefficients), two for each of the variables. The GMDH strategy allowed us to choose the most accurate model for each pair, most closely corresponding with the actual values of series one.

Then we calculated the values of each variable in the subsequent time point P_7 using the classical GMDH (formula (3) for x variable) for later comparison with the results of the modified method.

At the next step we found the values of the first and second order sensitivity functions at all the previously mentioned time points and we solved the system of linear (9), (10) and nonlinear (11), (12) equations in order to find the values of each relative increase for each process variable.

Both values of the sensitivity functions and relative increases could be interpolated, for example by the Lagrange's method. The obtained results were used to calculate the expected values of the variables of the process at subsequent time points. For the time series with visible regularities the charts of corresponding sensitivity functions clearly show their regularities (Fig. 2).

This paper presents the results of the prediction of z variable, because it appears to be more interesting due to the fact that the behavior of this variable is affected by both x and y variable. So we dealt with the partial functions $Z(z, x)$ and $Z(z, y)$.

We solved the system of equations (6) at time points $P_0 – P_6$. We obtained the coefficients $a_0 = –0.8174$, $a_1 = 2.9276$, $a_2 = 2.3909$, $a_3 = 0.0003$, $a_4 = –2.6732$, $a_5 = –0.5726$ for the partial function $Z(z, x)$. And $b_0 = –0.0843$, $b_1 = 2.2413$, $b_2 = 0.2659$, $b_3 = –0.0104$, $b_4 = –2.4462$, $b_5 = –0.2206$ for $Z(z, y)$.

We calculated the values of the first and second order sensitivity functions at each given time point. The sensitivity is a measure of the effect how on the changes of one factor affects another factor. The functions and approximate the course of the z values very well (as they tell us, what impact the z value has on itself). It can be seen that the values of the second order sensitivity functions and take the significantly different values. It is caused by differing values and signs of the coefficients $a_3 = 0.0003$ and $b_3 = –0.0104$. While the other corresponding values take significant absolute values. They tell us about the quantity of the impact of the z variable changes on the x and y values at subsequent time points.

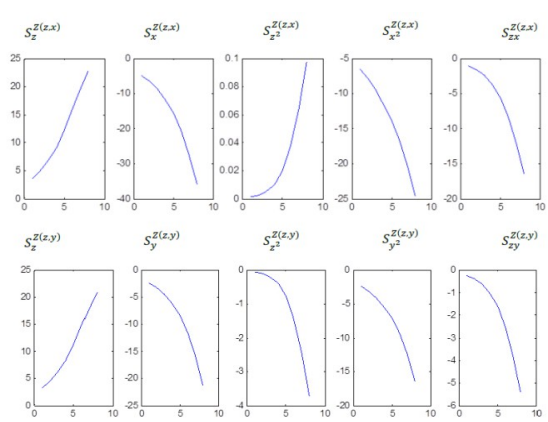

Fig. 2. Charts of the sensitivity functions for $Z(z, x)$ and $Z(z, y)$ models in $P_0 – P_7$ time points
Rys. 2. Wykresy funkcji czułości dla modeli $Z(z, x)$ oraz $Z(z, y)$ w punktach czasowych $P_0 – P_7$.

Next, we solved the system of linear and nonlinear equations using the previously calculated coefficients in order to find the values of relative increases for each model. Then we extrapolated the obtained values using the Lagrange's method for 1-3 steps forward at the most. After this, we had all necessary data to calculate the values of variables at P_7 time point (if need be at P_8 and P_9) using (9) and (10) with various

methods of solving nonlinear systems of equations. Then we used (3) and compared the calculations. The calculation results are given in Tab. 2. The most precise results of the forecast are marked in green (in **[7]** and **[3]**).

Tab. 2. The table shows the calculation results obtained for various methods. The time points P_7 – P_9 are predicted values for various models. **[2]** – The actual values of z variable of time series, **[3]** – The values of z obtained using the simplify model $Z(z, x)$ with the linear system of equation, **[4]** - The values of z obtained using the simplify model $Z(z, y)$ with the linear system of equations, **[5]** - The values of z obtained using the more precise model $Z(z, x)$ with the non-linear system of equations, **[6]** - The values of z obtained using the more precise model $Z(z, y)$ with the non-linear system of equations, **[7]** - The values of z obtained using the more precise model $Z(z, x)$ with the non-linear system of equations by the Newton-Ralphson method, **[8]** - The values of z obtained using the more precise model $Z(z, y)$ with the non-linear system of equations by the Newton-Ralphson method, **[9]** - The values of z obtained using the GMDH for $Z(z, x)$, **[10]** - The values of z obtained using the GMDH for $Z(z, y)$

Tab. 2. Tabela przedstawia wyniki obliczeń uzyskanych za pomocą różnych metod. Punkty czasowe P_7 – P_9 są wynikami predykcji. **[2]** – rzeczywiste wartości zmiennej z szeregu czasowego, **[3]** – wartości zmiennej z uzyskane z użyciem uproszczonego modelu $Z(z, x)$ z liniowym układem równań, **[4]** – wartości zmiennej z uzyskane z użyciem uproszczonego modelu $Z(z, y)$ z liniowym układem równań, **[5]** – wartości zmiennej z uzyskane z użyciem bardziej precyzyjnego modelu $Z(z, x)$ z nieliniowym układem równań, **[6]** – wartości zmiennej z uzyskane z użyciem bardziej precyzyjnego modelu $Z(z, y)$ z nieliniowym układem równań, **[7]** – wartości zmiennej z uzyskane z użyciem bardziej precyzyjnego modelu $Z(z, x)$ z nieliniowym układem równań rozwiązywanych metodą Newtona-Ralphsona, **[8]** – wartości zmiennej z uzyskane z użyciem bardziej precyzyjnego modelu $Z(z, y)$ z nieliniowym układem równań rozwiązywanych metodą Newtona-Ralphsona, **[9]** – wartości zmiennej z uzyskane z użyciem GMDH dla $Z(z, x)$, **[10]** – wartości zmiennej z uzyskane z użyciem GMDH dla $Z(z, y)$.

[1]	[2]	[3]	[4]	[5]
P_7	13,40973	13,41712	13,47693	13,43861
P_8	15,02293	15,02821	15,36752	15,18853
P_9	14,51877	14,25065	15,21499	14,8979

[6]	[7]	[8]	[9]	[10]
13,50794	13,41692	13,48183	13,39462	13,40137
15,60109	15,02822	15,40015	14,84608	14,91338
16,2043	14,26165	15,33063	13,43566	13,75658

Checking the accuracy of each of the applied at time points P_1 – P_6 models we used the least square method. It always gives us information about the smallest sum of the squared errors. There is no guarantee, however, that this result has any practical sense. In particular, if there is a lot of detached data, the results may have nothing to do with the actual trend line or the relationship between the phenomena described by the variables.

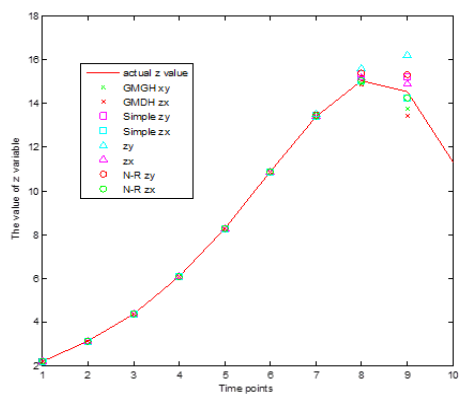

Fig. 3. The prediction results obtained using several models for three steps forward. P_7 – P_9– prediction.

Rys. 3. Wyniki predykcji na trzy kroki wprzód uzyskane za pomocą kilku różnych metod. Predykcja – P_7 – P_9.

The LSM values of the individual errors are as follows:

$LSM_{zx} = 1.5389 \cdot 10^{-7}$; $LSM_{zx\ simple} = 3.4633 \cdot 10^{-14}$;
$LSM_{zy} = 5.3951 \cdot 10^{-12}$; $LSM_{zx\ simple} = 2.212 \cdot 10^{-14}$;
$LSM_{zxN\text{-}R} = 9.9301 \cdot 10^{-16}$; $LSM_{GMDHzx} = 2.0788 \cdot 10^{-12}$;
$LSM_{zyN\text{-}R} = 1.3323 \cdot 10^{-15}$; $LSM_{GMDHzy} = 5.7665 \cdot 10^{-13}$.

The model with the partial function $Z(z, x)$ where the increases were obtained by solving the nonlinear system of equations using the Newton-Ralphson method turned out to be the most accurate: $LSM_{zxN\text{-}R} = 9.9301 \cdot 10^{-16}$ (the green circle in Fig. 3). It also gave the best results of the prediction at time point P_7 and even P_8 and P_9. The other results were also satisfactory. There was no need to apply the Kolmogorov-Gabor polynomial of higher than the second degree in order to continue the prediction at the next time point. So we chose the values obtained by the model, which proved to be the most accurate in the previous step from all which were tested, and repeated the procedure at new time points P_2 – P_8. The results are given in Tab. 3.

Tab. 3. The table shows the calculation results obtained in the second step of prediction. Denotations like in Tab. 2

Tab. 3. Tabela przedstawia wyniki obliczeń uzyskanych w drugim kroku predykcji. Oznaczenia jak w Tab. 2

[1]	[2]	[3]	[4]	[5]
P_8	15,02293	15,70352	15,71403	15,32028
P_9	14,51877	17,04771	17,15649	14,00139
P_{10}	11,28116	17,83075	18,5815	4,444241

[6]	[7]	[8]	[9]	[10]
15,49856	15,27	15,4871	15,82646	15,82556
15,36136	12,35539	14,86658	17,72176	17,68071
10,09401	-10,8398	5,245265	19,37835	19,10315

At the next step, with values taken from column **[5]** in Tab. 3 the prediction results are also satisfactory for our models. And at another step the results obtained by the classical GMDH turned out to be more accurate than those of the modified method.

Further results started to deviate significantly from the actual value of the *z* variable.

4. Conclusions

The above analysis and [8] show that use of sensitivity functions in solving the inverse problem of control is a legitimate strategy for prediction of the time series with significant regularities. The prediction results for a small number of the data sample proved to be very accurate and in the vast majority of cases was even more accurate than those obtained by the classical GMDH.

The sensitivity functions show the expected trend of the values of variables as well as the relationships between them. The data carried by them contain valuable information necessary to determine the exact variable which has a greater impact on the value of other variables and which of the tested models are better to estimate the future values of the time series.

The subject of our further research will be to identify the classes of functions for which the strategy of the sensitivity functions can give more accurate results than the classical GMDH.

5. References

[1] Zeliaś A.: Metody statystyczne, PWE, Warszawa 2000.
[2] Wentzell A. D.: Wykłady z teorii procesów stochastycznych, PWN, Warszawa 1980.
[3] Franklin G.F., Powell J.D.: Feedback Control of Dynamic Systems, Prentice Hall, Upper Saddle River, NJ, 5th edition, 2005.
[4] Farlow S.J.: Self-Organizing Methods in Modeling: GMDH Type Algorithms.: Marcel Decker Inc., New-York, 1984.
[5] Wiliński A.: GMDH – Metoda grupowania argumentów w zadaniach zautomatyzowanej predykcji zachowań rynków finansowych. Warszawa-Szczecin: Instytut Badań Systemowych Polskiej Akademii Nauk, 2009.
[6] Banks, H. T., Dediu S., Ernstberger S. L.: Sensitivity Functions and Their Uses in Inverse Problems, Journal of Inverse and Ill-posed Problems, Vol. 15, 7 (2007), p. 683–708.
[7] Findeisen W.: Struktury sterowania dla złożonych systemów, Oficyna Wydawnicza Politechniki Warszawskiej, Warszawa 1997.
[8] Thomaseth K., Cobelli C.: Generalized sensitivity functions in physiological system identification., Ann. Biomed. Eng., Vol. 27, 5 (1999), p. 607-616.

otrzymano / received: 21.11.2012
przyjęto do druku / accepted: 03.06.2013 *artykuł recenzowany / revised paper*