

3.4 Time series analysis and prediction by means of inductive models

Short-Term Processes Forecasting by Analogues Complexing GMDH Algorithm

Gregory Ivahnenko

Group MB, 10 Dallington Street, London, EC1V 0DB, United Kingdom

gmdhmail@gmail.com

Abstract. *In the report is described theoretical and practical results of complex systems forecasting by Analogues Complexing algorithm in case of short data samples. Described structure and modifications of the algorithm. Shown that complex application of inductive parametric, non-parametric and data mining methods allows to make all-round analysis of the object, investigate relationships of variables and simulate future development of processes.*

Keywords

Inductive modeling, Forecasting, Clusterization, Data mining, Analogues complexing, Decision support.

1 Introduction

The Group Method of Data Handling (GMDH) have some diversity of possibilities for all stages of modelling in comparison with another methods. It includes not only different models generators, criteria and models classes, but also consists of different kinds of algorithms which can be applied at all stages of data mining process in accordance to current problems [1]. Their solution require complex application of both non-parametric and parametric GMDH algorithms to get knowledge from data.

In the report is considered further investigation of non-parametric Analogues Complexing (AC) algorithm in connection with parametric algorithms for solution of forecasting problems in case of short or noised data.

2 Analogues Complexing GMDH algorithm

The algorithm is used for classification, clusterization problems solution and stepwise forecasting of multidimensional random processes by complexing (weighted addition) of analogues (similar patterns) taken from historical data. The main parameters of algorithm are optimized by sorting of discrete number of possible variants by inductive algorithm.

According to Ashby [2] diversity of control system or model must be not less than diversity of an object itself. The law of adequateness, given by S.Beer [3] establishes that for optimal control, objects should be compensated by corresponding black boxes of the control system. The only information about these boxes is that they have limited values of output variables, which are similar to the corresponding states of object. The equal fuzziness of the model and object is reached automatically if the object itself is used for forecasting. This is done by searching analogues from the given data sample which are equivalent to the physical model. Forecasts are not calculated in the classical sense but extracted from the table of observation data.

The main assumptions here are following:

- Investigated object can be described by multidimensional process;
- Multidimensional process is sufficiently representative, i.e. essential system variables are included into the input data sample and it contain sufficient number of observations;
- Part of previous behaviour of system in past can be repeated in future.

The Analogues Complexing method becomes effective in case when data is noised (fuzzy) or short. In this conditions application of usual data mining algorithms based on regression analysis is not possible [4].

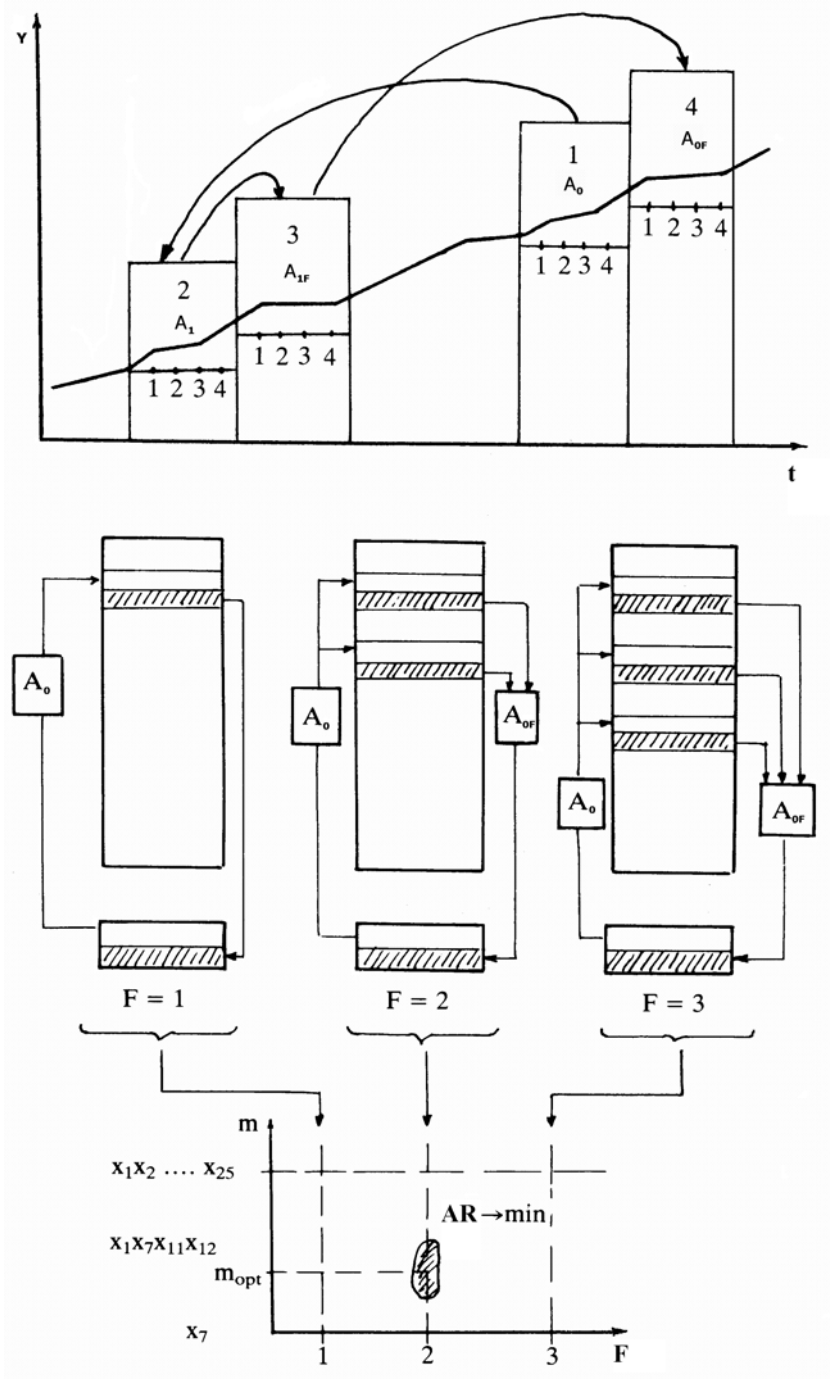


Fig 1. Analogues complexing and parameters optimization in AC algorithm.

If we succeed in search for the last part of behaviour trajectory (starting pattern 1), one or more analogous parts in the past (analogous pattern 2) the prediction 4 can be achieved by applying the known continuation of analogous patterns 3 (Fig.1). The pattern A_1 nearest to each given output pattern A_0 is called as its first analogue. The closest pattern A_2 is called as second analogue and so on. The pattern, which follows the first analogue in time, A_{1F} is called as its first analogues. The pattern, which follows the second pattern in time A_{2F} , is called as second analogue and so on. Patterns A_{iF} which are located after analogues A_i in time are called as their predictions (Fig.1). Forecast is calculated by complexing of optimal number of analogues.

During rigid complexing of F predictions by analogues, the prediction A_{0F} is defined using weights λ_i of analogues complexing :

$$A_{0F} = \sum_{i=1}^F \lambda_i A_{iF}, \quad (1)$$

$$\lambda_i = \frac{l_{0i}}{\sum_{i=1}^F l_{0i}}, \quad \sum_{i=1}^F \lambda_i = 1. \quad (2)$$

where l_{0i} – Euclid distance between initial pattern A_0 and analogues A_i ;

F – number of predictions.

During soft complexing of predictions by analogues the weights coefficients λ_i are defined by described rigid formulae (6) and then are adapted by sorting of their discrete values by inductive parametric algorithm.

The general problem of optimization of algorithm parameters should be solved in four dimensional space of sorting. There is needed to find out optimal, by regularity criterion $AR(s)$, values for such four parameters:

- set of input variables X ;
- number of analogues F for complexing;
- length of analogues k ;
- weight coefficients λ_i values for analogues complexing.

But it was founded that mutual relations between this parameters are such, that four dimensional sorting can be reduced to two one-dimensional (on X and λ_i) and one two-dimensional sorting (on F and k).

A. Search of optimal variables set.

At first is conducted one-dimensional sorting of input variables sets X while parameters F , k , and λ_i are fixed. This is strong (but not obligatory) optimization step of the algorithm. Optimal set of variables X is founded by one of inductive parametric algorithms.

B. Optimization of number and length of patterns.

Two-dimensional sorting for number of analogues F and their length k is important step of optimization. If necessary to reduce computations it can be simplified also to one-dimensional sorting for F by fixed length k . The matrixes of Euclid distances between analogues should be recalculated before sorting for each value of k .

According to [5] each pattern can be standardised additionally or a trend can be extracted. But such transformation can decrease accuracy of forecasting.

Preliminary clusterization of patterns and search of analogues only among the most nearest clusters to pattern A_0 had improved the results of forecasting. The Objective Computer Clusterization (OCC) GMDH algorithm is used to divide patterns into clusters according to minimal Euclidean distance. The output pattern A_0 is associated with the nearest cluster and search for analogues A_i is provided in this cluster only. Such additional step help to understand typical behaviours of the object and make interpretation of results much easier.

C. Evaluation of weight coefficients.

During investigation were tested several ways to determine values λ_i . The weights can be founded by rigid or by soft ways of complexing. Results obtained by the last way appear to be more accurate. Additional definition of weight coefficients by LSE method can be done in different ways [6].

During forecasting of very short time series the main problem was different length of them in the input data. To make possible complexing, when some analogues does not contain data, was applied procedure for correction of complexing weights for the last points of prediction.

D. Complexing (combining) of forecasts.

Each selected analogue A_i has its continuation in time A_{iF} which gives forecast A_{0F} . In such way we receive F forecasts needed for complexing. In literature [5] are described several ways to unite forecasts. In this algorithm unknown prediction A_{0F} is founded according to (6).

The GMDH algorithm of analogues complexing has advantage when number of input variables is greater than number of observations and for data with large dispersion of noises (for example stock market series).

The AC algorithm can be used not only for forecasting, but for clusterization and classification problems solution.

The next advantage of the AC algorithm over regression methods is that it can select a long-term forecast when number of observations is too small or less than number of variables. If also take assumption that founded analogues set define forecast on full length of input time series, than future values of forecast can be founded not only on one step ahead (as for regression models), but for the whole length of known analogues. This feature was used for search of forecast of launches of new products with very small number of observations and described below.

3 Example of long-term forecasting of product launch.

Modern analysis of marketing data is based mainly on application of regression models. Advanced systems for simulation of different scenarios were created. Handling of modern data bases require application of new modern methods which can detect interrelations and analyse influence of external factors on market development. Inductive algorithms can be such effective tool for discovery of relationships in data and rules of sales. The AC algorithm can be used when only a few observations are received. Such express-analysis of data is important for decision support and planning.

For forecasting of launches of some product categories was done complex analysis by developed software system. For this investigation was extracted data sample which contain variables for 186 shampoos, conditioners and series for total categories of this products. Each product is described by set of 14 variables including distribution, volumes and advertising features. As output the volume of sales was taken. It was needed to make express-analysis of new conditioner launch using 3 points of weekly observations using AC algorithm.

By Combinatorial GMDH algorithm we found out optimal input variables set on historical data for this category of products. Because the length of pattern k can be equal to three observations only the two-dimensional sorting was limited to number of pattern F selection. It was founded that for $F=2$ the regularity criterion $AR(s)$ value is minimal.

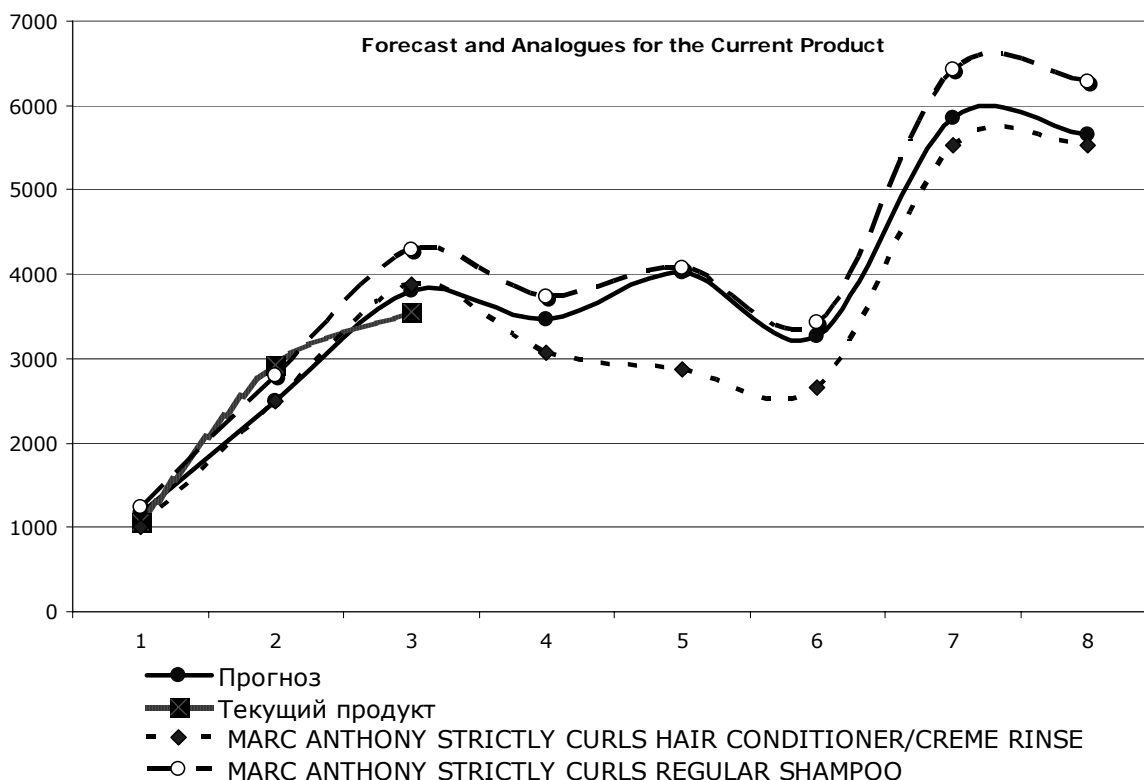


Fig. 2. Forecast for current product by AC algorithm.

The whole set of products was divided by OCC algorithm into 3 clusters. By minimum of Euclidian distance the current product was associated to the nearest second cluster and then were founded two nearest analogues in this cluster. For output variables of this two analogues the complexing was made to find out forecast (Fig. 2). The known values of current product were not used during complexing. The percentage error of forecast was MAPE = 10.8%.

All computations were made inside developed software system in very short time. After forecasting was produced the full report with initial data, series of errors, residuals, criteria values and models structures. On the separate sheet is possible to make simulations of possible developments in products sales, to interpret and investigate variables elasticity.

4 Conclusion

In the report is discussed application of GMDH method for forecasting of short data samples. The Analogues Complexing method was proven to be effective for forecasting and clusterization of data samples in many applications. It can be used for comprehensive analysis of the object together with another inductive parametric, non-parametric and data mining methods to get full picture of the process.

References

- [1] Ивахненко А.Г., Степашко В.С. Помехоустойчивость моделирования. К.: Наукова думка, 215 с. 1985. – <http://articles.gmdh.net>.
- [2] Ashby D. An introduction to cybernetics. J. Wiley, New York, 1958.
- [3] Beer S. Cybernetics and Management, English Univ.Press, London, 1959. – 280с.
- [4] Madala H.R. and Ivakhnenko A.G. Inductive Learning Algorithms for Complex Systems Modeling, CRC Press Inc., 1994, p.384.
- [5] Mueller J.-A., Lemke F. Self-Organizing Data Mining. Extracting Knowledge From Data. Trafford Publishing, Canada, 2003. – <http://knowledgeminer.com>.
- [6] Ивахненко А.Г., Богаченко Н.Н., Ли Тянь Мин. Безмодельное прогнозирование случайных процессов при помощи комплексирования прогнозов по аналогам. Проблемы управления и информатики. №4: 111-188, 1997.