Inductive Self-Organising Algorithm for Maximum Electrical Load Prediction

Ivakhnenko, G.A.
National Institute for Strategic Studies, vul. Pirogova 7-a, Kyiv 01030 Ukraina
e-mail: Gai@niss.gov.ua http://come.to/GMDH

Ivakhnenko, A.G.
International Centre of Informational Technologies and Systems of the National Ac.Sci., Ukraina, Kyiv
e-mail: gai@gmdh.kiev.ua

Report

Abstract: *This article presents a self-organising Combinatorial GMDH algorithm that provides sorting of linear and non-linear polynomial models. The calculated data shows that the proposed method is able to select simple models characterized by a good prediction ability and thus provide a considerable interest in the processes prediction.*

1. Introduction

Decision making in such areas as process analysis in electroenergy sector, macroeconomy, financial forecasting, analysis and another requires tools, which are able to get accurate models on basis of processes forecasting. Such objects are complex ill-defined systems that can be characterised by:

- inadequate a priori information;
- great number of immeasurable variables;
- noisy and extremely short data samples;
- ill-defined objects with fuzzy characteristics.

Problems of complex objects modelling such as analysis and prediction of maximum electrical load and other, cannot be solved by deductive logical-mathematical methods with needed accuracy. In this case knowledge extraction from data, i.e. to derive a model from experimental measurements, has advantages in cases of rather complex objects, being only little a priori knowledge or no definite theory particularly for objects with fuzzy characteristics on hand. This is especially true for objects with fuzzy characteristics.

GMDH type neural networks can overcome these problems - it can pick out knowledge about object directly from data sampling. The Group Method of Data Handling (GMDH) is the inductive sorting-out method, which has advantages in the cases of rather complex objects, having no definite theory, particularly for the objects with fuzzy characteristics.

2. Group Method of Data Handling (GMDH)

2.1. Brief description

The Group Method of Data Handling (GMDH) is self-organizing approach based on sorting-out of gradually complicated models and evaluation of them by external criterion on separate part of data sample.

Inductive GMDH algorithms gives possibility to find automatically interrelations in data, select optimal structure of model or network and increase the accuracy of existing algorithms. As input variables can be used any parameters, which can influence on the process. Linear or non-linear, probabilistic models or clusterizations are selected by minimal value of an external criterion. GMDH algorithms are rather simple and they get information directly from data sample.

This self-organizing approach is different from deductive methods or networks used commonly for modelling on principle. It has inductive nature - problems solution is based on sorting procedure by external criterion. The effective input variables, number of layers and neurons in hidden layers, optimal model structure are determined automatically. This is based on that fact that external criterion characteristic have minimum during complication of model structure. It was proved, that for inaccurate, noisy or small data can be found best optimal simplified model, accuracy of which is higher and structure is simpler than structure of usual full physical model. For real problems with noised or short data samples, simplified forecast models becomes more effective.

Group Method of Data Handling was applied in many countries for data mining and knowledge discovery, forecasting and systems modelling, optimization and pattern recognition. Since 1968 many books, more than 230 doctoral dissertations were devoted to investigations in very different fields. The GMDH theory and source code of some algorithms was also published in [1,2,3] and at GMDH website [http://GMDH.come.to]

GMDH solves, by sorting-out procedure, the multidimensional problem of model optimization:

$$\tilde{g} = \arg \min_{g \subset G} CR(g), \; CR(g) = f\,(P, S, \xi^2, T, V) \qquad (1)$$

where: G - set of considered models; CR is an external criterion of model g quality from this set; P - number of variables set; S - model complexity; ξ^2 - noise dispersion; T - number of data sample transformation; V - type number of reference function. For definite reference function, each set of variables corresponds to definite model structure $P = S$. Problem transforms to much simpler one-dimensional

$$CR(g) = f\,(S),$$

when ξ^2= const, T = const, and V = const.

Method is based on the sorting-out procedure, i.e. consequent testing of models, chosen from set of models-candidates in accordance with the given criterion. Most of GMDH algorithms use the polynomial reference functions. General connection between input and output variables can be expressed by Volterra functional series, discrete analogue of which is Kolmogorov-Gabor polynomial [1]:

$$y = a_0 + \sum_{i=1}^{M} a_i x_i + \sum_{i=1}^{M}\sum_{j=1}^{M} a_{ij} x_i x_j + \sum_{i=1}^{M}\sum_{j=1}^{M}\sum_{k=1}^{M} a_{ijk} x_i x_j x_k,$$

where $X(x_1, x_2, \ldots, x_M)$ - input variables vector;

$A(a_1, a_2, \ldots, a_M)$ - vector of coefficients or weights.

Components of the input vector X can be independent variables, functional forms or finite difference terms. Other non-linear reference functions, such as difference, logistic, harmonic can also be used for model construction. The method allows to find simultaneously the structure of model and the dependence of modelled system output on the values of most significant inputs of the system.

The GMDH theory solve the problems of:

- long-term forecasting;
- short-term forecasting of processes and events;
- identification of physical regularities;
- approximation of multivariate processes;
- physical fields extrapolation;
- data samplings clusterization;
- pattern recognition in the case of continuous-valued or discrete variables;
- diagnostics and recognition by probabilistic sorting-out algorithms;
- vector process normative forecasting;
- modeless processes forecasting using analogues complexing;
- self-organization of twice-multilayered neuronet with active neurones.

In [2] were obtained the theoretical grounds of GMDH effectiveness as adequate method of robust forecasting models construction. Essence of it consists of automatically generation of models in given class by sequential selection of the best of them by criteria, which implicitly by sample dividing take into account the level of indeterminacy.

Since 1967 a big number of GMDH technique implementations for modelling of economic, ecological, environmental, medical, physical and military objects were done in several countries. Some approaches are used in USA by Megaputer Corp. in “PolyAnalyst”, Ward Systems Group, Inc. in "NeuroShell2", AbTech Corp. "ModelQuest", Barron Associates Co. "ASPN", and DeltaDesign Berlin Software "KnowledgeMiner" commercial software tools.

There is a large spectrum of GMDH algorithms. It includes supervised (combinatorial algorithm, iteration algorithm, objective system analysis, harmonical, etc.), and unsupervised (objective computer clusterization, analogues complexing) methods (see for review [4). The choice of the appropriate GMDH algorithm depends on the specificity of the problem to be solved. The specificity of the maximum electrical load prediction tasks can be summarized as follows: there is a small number of input variables, some of these variables are irrelevant and not correlated, and only restricted input data set is available. The GMDH approach is well suited to solve such problems.

Recent developments of the GMDH have led to neuronets with active neurons, which realise twice-multilayered structure: neurons are multilayered and they are connected into multilayered structure. This gives possibility to optimise the set of input variables at each layer, while the accuracy increases. Not only GMDH algorithms, but also many modelling or pattern recognition algorithms can be used as active neurons. Its accuracy can be increased in two ways:

- each output of algorithm (active neuron) generate new variable which can be used as a new factor in next layers of neuronet;
- the set of factors can be optimised at each layer. The factors (including new generated) can be ranked after their efficiency and several of the most efficient factors can be used as inputs for next layers of neurons. In usual once-multilayered ANN the set of input variables can be chosen once only.

2.1. The Combinatorial GMDH algorithm (COMBI)

The flowchart of the algorithm is shown in Fig. 1. The input data sample is a matrix containing N levels (points) of observations over a set of *M* variables. The sample is divided into two parts. Approximately two-thirds of points make up the learning subsample N_A, and the remaining one-third of points (e.g. every third point) with same variance form the check subsample N_B. Before dividing, points are ranged by variation value. The learning sample is used to derive estimates for the coefficients of the polynomial, and the check subsample is used to choose the structure of the optimal model, that is, one for which the external regularity criterion *AR(s)* takes on a minimal value:

$$AR(s) = \frac{1}{N_B} \sum_{i=1}^{N_B} (y_i - y_i(B))^2 \to \min \qquad (2)$$

or better to use the cross-validation criterion *PRR(s)* (it takes into account all information in data sample and it can be computed without recalculating of system for each checking point):

$$PRR(s) = \frac{1}{N} \sum_{1}^{N} \left[y_i - y_i(B) \right]^2 \to \min, \qquad N_A = N - 1; \quad N_B = 1.$$

To test a model for compliance with the differential balance criterion, the input data sample is divided into two equal parts. The criterion requires to choose a model that would, as far as possible, be the same on both subsamples. The balance criterion will yield the only optimal physical model solely if the input data are noisy.

To obtain a smooth exhaustive-search curve, which would permit one to formulate the exhaustive-search termination rule, the exhaustive search is performed on models classed into groups of an equal complexity. For example, the first layer can use the information contained in every column of the sample; that is full search is applied to all possible models of the form:

$$y = a_0 + a_1 x_i, \qquad i = 1,2,...,M \, . \qquad (3)$$

Non-linear members can be taken as new input variables in data sampling. The output variable is specified there in advance by the experimenter. At next layer are sorted all models of the form:

$$y = a_0 + a_1 x_i + a_2 x_j, \qquad j = 1,2,...,M \qquad (4)$$

The models are evaluated for compliance with the criterion, and so on until the criterion value decrease. For limitation of calculation time recently it was proposed during full sorting of models to range variables according to criterion value after some time of calculation or after some layers of iteration. Then full sorting procedure continues for selected set of best variables till the minimal value of criterion will be found. This gives possibility to set much more input variables at input and to save effective variables between layers to found optimal model.

A salient feature of the GMDH algorithms is that, when they are presented continuous or noisy input data, they will yield as optimal some *simplified non-physical model*. If is only in the case of discrete or exact data that the exhaustive search for compliance with the precision criterion will yield what is called a *physical model*, the simplest of all unbiased models. With noisy and continuous input data, simplified (Shannon) models prove more precise [2,4] n approximation and for forecasting tasks.

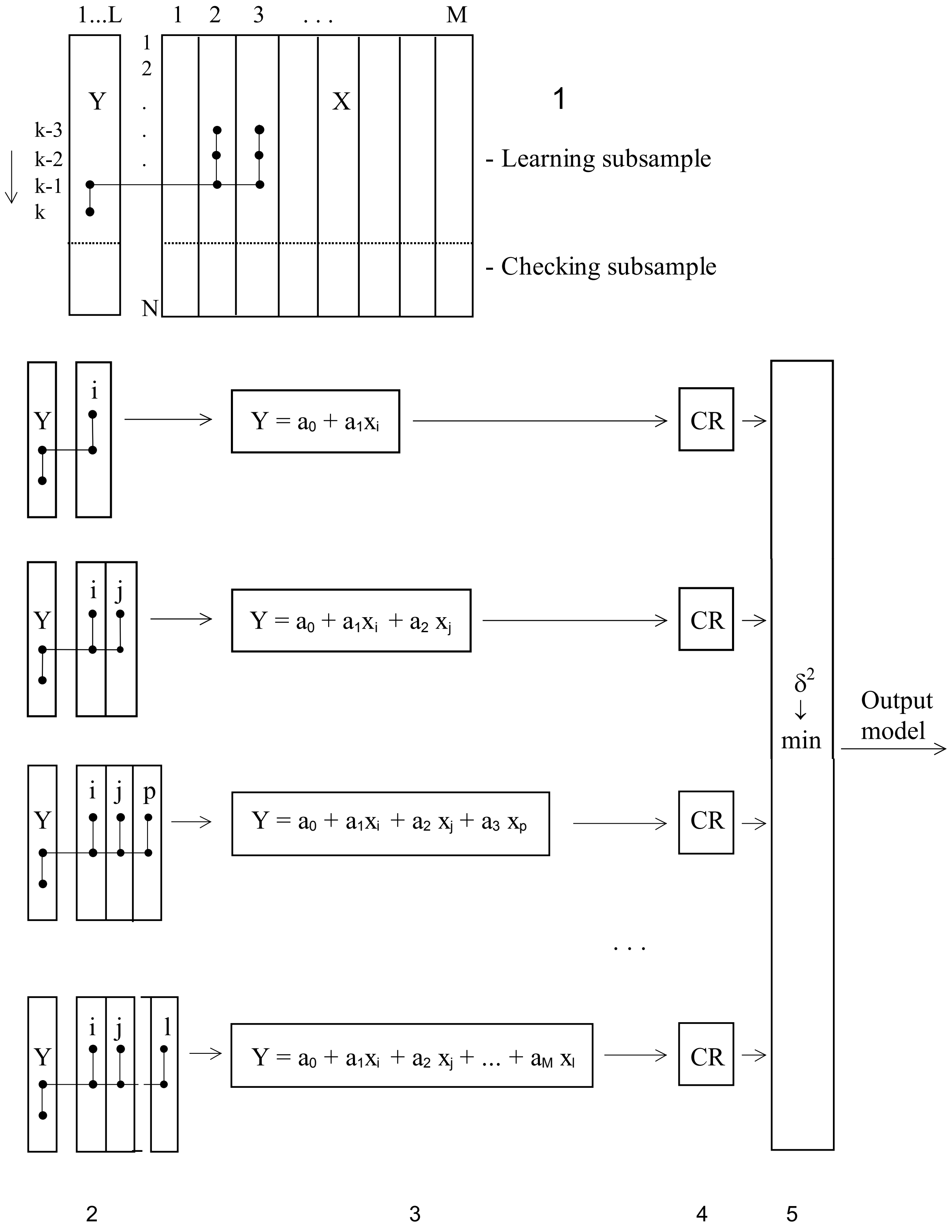

Fig. 1. Combinatorial GMDH algorithm.

1 - data sampling;
2 - layers of partial descriptions complexing;
3 - form of partial descriptions;
4 - choice of optimal models;
5 - additional model definition by discriminating criterion.

3. Experimental results.

The input data set for maximum electrical load prediction for competition within the EUNITE network consists of three variables: electrical load, average daily temperatures and the occurrence of holidays in the period from 1997 – 1998.

At first stage of investigation several secondary variables were generated. As additional inputs results of ARIMA model, exponential smoothing, seasonal decomposition were also used. During preliminary parameter selection the Combinatorial GMDH algorithm show us what variables better to use – it were selected according to value of external criterion. In result the data sample contain the following variables:

X_1 -daily average electrical load
X_2 – correlation with previous day of electrical load
X_3 – covariation
X_4 – standardized maximum electrical load
X_5 - 3-points moving average
X_6 - autocorrelation
X_7 - residualizing
X_8 - differencing
X_9 - trend subtract
X_{10} - 4253H filter

Results of seasonal decomposition:

X_{11} - moving averages
X_{12} - seasonal factors
X_{13} - adjusted series
X_{14} - smoothed trend cycle
X_{15} – irregular component

Results of exponential smoothing:

X_{16} - smoothed multiplicative model
X_{17} - smoothed additive model
X_{18} - 1/Y
X_{19} - Y^2
X_{20} - ln(Y)
X_{21} – exponentially smoothed Y (a=0.2)
$X_{23} = Y_{(t-1)}$
$X_{23} = Y_{(t-2)}$
$X_{24} = Y_{(t-3)}$
$X_{25} = Y_{(t-4)}$
$X_{26} = Y_{(t-5)}$
$X_{27} = Y_{(t-6)}$
$X_{28} = Y_{(t-7)}$
$X_{29} = Y_{(t-14)}$

Output variable:

Y - daily maximum electrical load

During computation we have found that X_2, X_5, X_9, X_{12}, X_{19} variables are not effective. For each forecast value was found forecasting polynomial model by Combinatorial GMDH algorithm. The errors values, average for 31 models are following:

MSE	20.229
MAXIMAL	15.345
Correlation	0.975
R^2	0.951
MAPE	2.346%

4. Conclusions

The calculated data shows that the proposed method is able to select simple models characterized by a high prediction ability and thus provide a considerable interest in the processes prediction. Unfortunately analysis was rather rough and different other effective GMDH techniques (e.g. Harmonical GMDH algorithm, Twice-multilayered neural networks) which can increase significantly accuracy of forecast should be applied.

References

1. Mueller, J.-A., Lemke, F. Self-Organising Data Mining Libri, Hamburg, 2000, ISBN 3-89811-861-4, http://www.knowledgeminer.net
2. Madala,H.R. and Ivakhnenko,A.G. Inductive Learning Algorithms for Complex Systems Modeling. CRC Press Inc., Boca Raton, 1994, p.384.
3. Farlow,S.J.,(ed.) Self-organising Methods in Modeling (Statistics: Textbooks and Monographs, vol.54), Marcel Dekker Inc., New York and Basel, 1984.
4. http://GMDH.come.to