Robust Structural Modeling and Outlier Detection with GMDH-Type Polynomial Neural Networks

Tatyana Aksenova[1,2], Vladimir Volkovich[3], and Alessandro E.P. Villa[1]

[1] Inserm U318, Laboratory of Neurobiophysics, University Joseph Fourier, Grenoble, France
{Tatyana.Aksyonova, Alessandro.Villa}@ujf-grenoble.fr
http://www.nhrg.org/
[2] Institute of Applied System Analysis, Prospekt Peremogy, 37, Kyiv 03056, Ukraine
[3] International Researching-Training Center of Information Technologies, Glushkova 40, 252022, Kyiv, Ukraine
volk@volk.kiev.ua

Abstract. The paper presents a new version of a GMDH type algorithm able to perform an automatic model structure synthesis, robust model parameter estimation and model validation in presence of outliers. This algorithm allows controlling the complexity – number and maximal power of terms – in the models and provides stable results and computational efficiency. The performance of this algorithm is demonstrated on artificial and real data sets. As an example we present an application to the study of the association between clinical symptoms of Parkinsons disease and temporal patterns of neuronal activity recorded in the subthalamic nucleus of human patients.

1 Introduction

Artificial Neural Networks (ANN) have been successfully applied in many fields to model complex non–linear relationships. ANNs may be viewed as the universal approximators but the main disadvantage of this approach is that detected dependencies are hidden within the neural network structure. Conversely, Group Method of Data Handling (GMDH) [1] are aimed to identify the functional structure of a model hidden in the empirical data. The main idea of GMDH is the use of feedforward networks based on short-term polynomial transfer function whose coefficients are obtained using regression technique combined with the emulation of the self-organizing activity for the neural network (NN) structural learning. In order to reduce the sensitivity of GMDH to outliers a Robust Polynomial Neural Network (RPNN) approach was recently developed [2]. This paper presents a new version of RPNN using new robust criteria for model selection and measures of goodness of fit and a demonstration of its performance on artificial and real data sets.

W. Duch et al. (Eds.): ICANN 2005, LNCS 3697, pp. 881–886, 2005.

2 GMDH Approach

The GMDH approach for complex system modeling and identification is based on given multi-unit–single-output data invented by Ivakhnenko [1]. Traditional GMDH is a multi-layered perceptron type NN formed by neurons whose transfer function g, $g = a + bw_i + cw_j + dw_iw_j + ew_i^2 + fw_j^2$ is a short-term polynomial of two variables w_i, w_j. The GMDH training algorithm is based on an evolutionary principle. The algorithm begins with regression-type data, the observations of vector of independent variables $x = (x_1, \ldots, x_m)^T$ and one dependent variable y. The data set is subdivided into training and test sets. At the 1st layer all possible combinations of two inputs generate the first population of neurons according to the transfer function g. The size of the population at the 1st layer is equal to C_m^2. The coefficients of the polynomials of g are estimated by Least Square fitting using the training set. The best neurons are selected by evaluating the performance on the test set according to a criterion value. The outputs of selected neurons of the first layer are treated as the inputs to the neurons of the 2nd layer, and so on for the next layers. The size of the population of the successive layers become equal to C_f^2. The process is terminated if there is no improvement of the performance according to the criterion. The GMDH model can be computed by tracing back the path of the polynomials. The composition of quadratic polynomials of g forms a high-order regression polynomial known as the Ivakhnenko polynomial. Notice that the degree of the polynomial doubles at each layer and the number of terms in the polynomial increases.

3 Robust Polynomial Neural Networks

Basically RPNN are described as follows (see [2,3] for more details). Let $\mathbf{x} = (x_1, \ldots, x_m)^T$ be the vector of input variables and let y be the output variable that is a function of a subset of input variables $y = u(x_{i1}, x_{i2}, \ldots, x_{ip})$. Let $\mathbf{X} = (x_{ij})$ be a $[m \times n]$ matrix and $\mathbf{Y} = (y_1, \ldots, y_n)^T$ the vector of observations. The random errors ξ of observations are assumed to be uncorrelated, identically distributed with finite variance $\mathbf{Y} = E(\mathbf{Y}|\mathbf{X}) + \xi$. The goal of the method is to find a subset of variables $x_{i1}, \ldots, x_{ik}$ and a model belonging to the class of polynomial that minimizes some criteria values (CR). Thus, model identification means both structure synthesis and parameters estimation. The main modifications according to the original GMDH are the following:

1. An expanded vector of initial variables $\mathbf{x} = (x_1, \ldots, x_m, x_{m+1}, x_{m+2})^T$, $x_{m+1} = 1$, $x_{m+2} = 0$ is available at each layer;

2. The following nonlinear transfer function which generates the class of polynomials is used [3]:

$$g(w_i, w_j, w_k) = aw_i + bw_jw_k, \qquad i, j, k = 1 \ldots m \qquad . \tag{1}$$

Triplets of inputs are considered instead of pairs. The coincidences of indexes lead up to triple the number of connections. Neurons with one or two inputs as well as several transform functions (including the linear one) are generated

according to Eq. 1 and additional variables $x_{m+1} = 1$ and $x_{m+2} = 0$. Notice that only two coefficients a and b are estimated. In traditional GMDH the Mean Least Square (MLS) method is used. Thus the second order matrices are only inverted. This provides fast learning of NN. The number of neurons at each layer of the net that depends on the form of the transfer function g and the number f of output variables which were selected from previous layer equals C^3_{m+2+f} .

3. Each term $x_i^{q1}, \ldots, x_j^{q2}$ in the equation is coded as a product of the appropriate powers of a prime numbers, i.e. the polynomial is coded by a vector of Gedels numbers [3]. Because of the one-to-one correspondence between the terms of the polynomials and their Gedels numbers this coding scheme can be used to transfer the results of the ANN to the parametric form of equation.

4. The polynomials of high power are unstable and sensitive to outliers. Therefore, a twice-hierarchical ANN structure based on the polynomial complexity control [2] was proposed to increase the stability and computational efficiency of GMDH. This structure allows the convergence of the MLS coefficients, as proven mathematically for algorithm with linear transform [4]. The vector $(p, c)^T$, where c is the number of terms and p is the power of the polynomial is considered as the polynomial complexity. Gedels coding scheme allows to calculate the number of terms for each intermediate model that equals to the number of non zero element of its vector of Gedels numbers. The power of intermediate model $g(w_i, w_j, w_l) = aw_i + bw_jw_l$ is controlled by the condition $p(g(w_i, w_j, w_l)) = maxp(w_i), p(w_j), p(w_l)$ where $p(w_i), p(w_j), p(w_l)$ are the power of inputs w_i, w_j, w_l. This allows to control the complexity by restricting the class of the models by $p(w_i) < p_{max}$ and $c < c_{max}$. The RPNN are twice-multilayered since multilayered neurons are connected into a multilayered net. The *external iterative procedure* controls the complexity of the models, i.e. the number of the terms and the power of the polynomials in the intermediate models. The best models form the initial set for the next iterative procedure. The *internal iterative procedure* realizes a search for optimal models given the fixed complexity and discard models that are out of the specified range. Both external and internal iteration procedures are terminated if there is no improvement of the criterion values CR.

5. Robust M–estimates [5] of the coefficients a and b of the transfer functions $g(w_{j1}, w_{j2}, w_{j3}) = aw_{j1} + bw_{j2}w_{j3}$ were applied instead of MLS estimates.

6. Robust versions of CR are used for model selection:

$$CR1 = \frac{\hat{\sigma}}{n-p} \sum_{i=1}^{n} \rho(r_i/\hat{\sigma}) \quad , \quad CR2 = \hat{\sigma} \frac{n+p}{n-p} \sum_{i=1}^{n} \rho(r_i/\hat{\sigma}) \quad . \tag{2}$$

If the data is splitted into training and test sets A and B, then the robust version of regularity criterion AR used in GMDH [1] is implemented. The parameters $\hat{a}_A$, $\hat{b}_A$, and the variance $\hat{\sigma}_A$ estimated on the set A are used to calculate the residuals r_i for the set B. Then, the regularity criterion AR is expressed by $AR = \sigma_A^2 \sum_{i \in B} \rho(r_i/\hat{\sigma}_A)$.

7. The ρ-test (robust variant of F-criteria) and R_n^2-test [6] are applied to the final models for the canonical hypothesis $H_0 : \beta_i = 0$, β is the vector of

parameters of the resulted model, to avoid the appearance of spurious terms. Robust correlation and robust deviation for both training set A and test set B are used as measures of goodness of fit.

8. The residuals r_i, of the final models are used for the outlier detection: $outlier = 0, if|r_i| \leq k\hat{\sigma}$, otherwise $outlier = 1$.

4 Validation on an Artificial Data Set

Let us consider the vector of $m = 5$ input variables $\mathbf{x} = (x_1, \ldots, x_5)^T$, and the fourth power polynomial $y = 10.0 + 1.0 \cdot x_1 \cdot x_5^3 + \xi$ generally used for testing GMDH. The matrix $\mathbf{X} = (x_{ij})$ $[5 \times 15]$, $n = 15$ was generated at random with a uniform distribution on the interval $[1, 10]$. Random values ξ were generated according to the model of outliers $P_\delta(\xi) = (1 - \delta)\phi(\xi) + \delta h(\xi)$. Here $\phi(\xi)$ is the Normal distribution density $N(0, \sigma_b)$; $h(\xi)$ is the distribution density of the outliers $N(0, \sigma_{out})$; δ is the level of the outliers. Twenty realizations were considered for the following combinations of parameters: (A) $\sigma_b = 10$, $\delta = 0$; (B) $\sigma_b = 10$, $\delta = 0.2$, $\delta_{out} = 1000$; (C) $\sigma_b = 10$, $\delta = 0.2$, $\delta_{out} = 2000$; (D) $\sigma_b = 10$, $\delta = 0.2$, $\delta_{out} = 3000$. Structural indexes $StrInd$ were determined for each term of the equation: $StrInd = 1$ if the term was present in the synthesized equation and $StrInd = 0$ otherwise. The mean value of the structural indexes corresponds to the frequency of the appearance of the term in the resulted equations over all computational experiments. Table 1 shows that RPNN provides the best structural synthesis irrespective of the increasing variance of the outliers. Table 2 summarizes the results of the coefficients estimation with RPNN and PNN (MLS). Table 3 presents the quality of approximation with the measures of goodness of fit of the calculated model according to the exact model $y_{exact} = 10.0 + 1.0 \cdot x_1 \cdot x_5^3$, μ_{exact} is the mean value:

$$MSD = \frac{1}{n}\sum(y_{calc} - y_{exact})^2, R^2 = \frac{\sum(y_{exact} - \mu_{exact})^2 - \sum(y_{calc} - y_{exact})^2}{\sum(y_{exact} - \mu_{exact})^2}. \tag{3}$$

Table 1. Mean values of structural indexes $StrIn$ for the constant, monomial $x_1 \cdot x_5^3$ and additional terms with significant (signif.) and non-significant (n.s.) coefficients. δ: level of the outliers, σ_b, σ_{out}: : SD of basic Normal distribution and of the outliers.

		MLS		**RPNN**			**PNN with MLS**		
	δ	$\delta = 0$		$\delta = 0.2$			$\delta = 0.2$		
	σ_b	0	10	$\sigma_b = 10$			$\sigma_b = 10$		
	σ_{out}	$--$	$--$	1000	2000	3000	1000	2000	3000
	const	1.00	0.76	0.47	0.53	0.47	0.07	0.21	0.07
	$x_1x_5^3$	1.00	1.00	1.00	1.00	1.00	0.80	0.50	0.33
add	*n.s.*	0.00	0.65	0.60	0.53	0.60	0.40	0.64	0.33
	significant	0.00	0.00	0.00	0.00	0.00	1.33	0.71	1.60

Table 2. Coefficients estimated in case the terms were present in the resulted equation

		MLS		**RPNN**			**PNN with MLS**		
δ		$\delta = 0$		$\delta = 0.2$			$\delta = 0.2$		
σ_b		0	10	$\sigma_b = 10$			$\sigma_b = 10$		
σ_{out}		$--$	$--$	1000	2000	3000	1000	2000	3000
const	*mean*	10.0	11.69	10.86	10.54	10.53	687.00	127.50	247.29
	SD	–	2.59	1.96	1.90	1.88	–	1041.41	–
$x_1x_5^3$	*mean*	1.0	0.997	0.996	0.992	0.996	1.0541	1.1104	1.0673
	SD	–	0.004	0.003	0.003	0.003	0.1516	0.6058	0.2412
σ_b	*mean*	10.0	8.25	14.50	12.53	13.21	304.62	537.20	785219
	SD	–	1.99	6.93	4.90	5.64	197.60	261.58	758152.26

Table 3. The measures of goodness of fit according to the exact data

		MLS	**RPNN**			**PNN with MLS**		
δ		$\delta = 0$	$\delta = 0.2$			$\delta = 0.2$		
σ_b		$\sigma_b = 10$	$\sigma_b = 10$			$\sigma_b = 10$		
σ_{out}		$--$	1000	2000	3000	1000	2000	3000
MSD	*mean*	5.49	6.13	5.83	5.96	272.43	447.79	711.28
	SD	1.82	3.08	2.97	3.12	139.11	159.32	309.02
R^2	*mean*	0.99996	0.99995	0.99996	0.99995	0.91598	0.78919	0.58605
	SD	0.00002	0.00004	0.00004	0.00004	0.07346	0.16420	0.34052

5 Application to Experimental Data

RPNN was applied to study the association between clinical symptoms of Parkinsons disease and firing patterns in the subthalamic nucleus of patients that underwent surgical operation for Deep Brain Stimulation (DBS) [7]. The set of parameters determined from the neurological examination of the patients (x_i) are based on scores defined by the Unified Parkinsons Disease Rating Scale (x_1=RT: resting tremor; x_2=AT: action tremor, i.e. essential tremor during voluntary movement; x_3=RG: rigidity of upper limbs; x_4=AK: akinesia of upper limbs) and the parameters defining the firing activity (y_i) are obtained from the electrophysiological recordings (y_1=Syn: percentage of pairs of units with synchronous firing; y_2=Bst: % of units with bursting activity; y_3=(1-2): % of units with oscillatory activity [1-2 Hz]; y_4=(4-6): % of units with oscillatory activity [4-6 Hz]; y_5=(8-12): % of units with oscillatory activity [8-12 Hz]; y_6=FR: average firing rate in the subthalamic nucleus of patients operated at the Grenoble University Hospital. We have proceeded by considering the clinical parameter vector X as the independent variable and the neurophysiological vector Y as the dependent variable. However, the relation of causality between clinical and neurophysiological parameters is not known and we have analyzed the data considering also X dependent on Y. Two outliers exceeding 3σ-confidential interval were detected (Fig. 1a). The final result of the analysis allowed to generate a

new model of the associations between clinical symptoms and neurophysiological data (Fig. 1b). In this Figure the arrows show the presence of the variables in the models and the sign of the corresponding terms. Notice that only models with criteria values $R^2 > 0.6$ were considered.

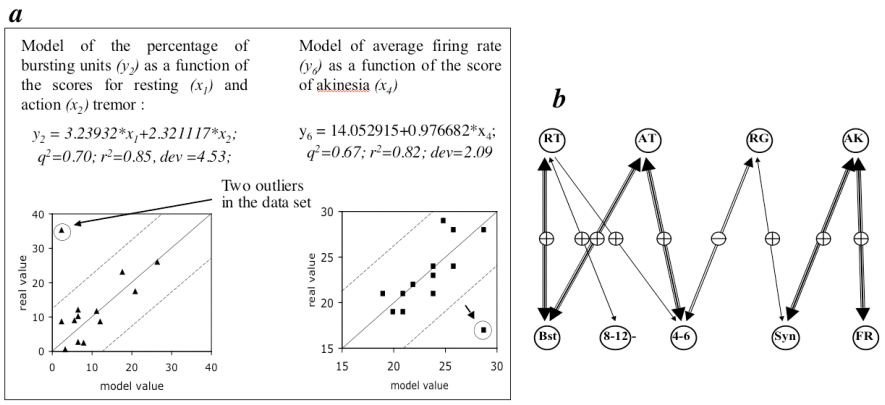

Fig. 1. (a) Examples of polynomial estimates; (b) Results of modeling presented as a scheme of the dependencies

Acknowledgments. The authors thank Pr. A.-L. Benabid, Dr. S. Chabardes and all members of the neurosurgical and neurological teams of Inserm U318 at Grenoble University Hospital for providing the data used for an application of this technique.

References

1. Madala, H., Ivakhnenko, A.: Inductive Learning Algorithms for Complex Systems Modeling. CRC Press Inc., Boca Raton, FL, USA (1994)
2. Aksenova, T.I., Volkovych, V., Tetko, I.: Robust Polynomial Neural Networks in Quantative-Structure Activity Relationship Studies. SAMS **43** (2003) 1331–1341
3. Yurachkovsky, Y.: Restoration of polynomial dependencies using self-organization. Soviet Automatic Control **14** (1981) 17–22
4. Yurachkovsky, Y.: Convergence of multilayer algorithms of the Group Method of Data Handling. Soviet Automatic Control **14** (1981) 29–35
5. Huber, P.: Robust Statistics. John Wiley & Sons Inc. (2003)
6. Hampel, F., Ronchetti, E.M., Rousseeuw, P., Stahel, W.: Robust Statistics: The Approach Based on Influence Function. John Wiley & Sons Inc. (2005)
7. Chibirova, O., Aksenova, T.I., Benabid, A., Chabardes, S., Larouche, S., Rouat, J., Villa, A.: Unsupervised Spike Sorting of extracellular electrophysiological recording in subthalamic nucleus parkinsonian patients. Biosystems **79** (2005) 59–71