Time Series Prediction by means of GMDH Analogues Complexing and GAME

Josef Bouška, Pavel Kordík

Dept. of Computer Science and Engineering, Karlovo nam. 13, 121 35 Praha 2, Czech Rep.

{bouskj1,kordikp}@fel.cvut.cz

Abstract. *For time series prediction we can use either parametric or nonparametric models. In this paper we study properties of both approaches for short and medium term prediction intervals. We compare the accuracy of GMDH Analogues Complexing as typical nonparametric method and the Group of Adaptive Models Evolution (GAME) as a parametric method. In our study, we focus on medical data from Motol hospital in Prague and horticulture data from Hort Research New Zealand.*

Keywords

Inductive modeling,
Analogues Complexing GMDH,
GAME,
Time series prediction.

1 Introduction

Prediction of time series determines future values based on measuring previous values of that series. The goal is to predict unknown future values from available data. There are many methods for prediction time series, ranging from statistical methods to neural networks as typical black box methods. Many of these methods are parametric, meaning that some parameters are being adjusted to fit the time series and to estimate future values.

In this article we focus on two different methods based on inductive modeling. As a nonparametric method we describe our implementation of well known GMDH Analogues Complexing (AC) inductive method [1, 3, 4].

We compare the performance of the AC method with the performance of the GAME method [5] which is parametric. The comparison is performed on real data sets, first on horticulture data (water consumption of mandarin tree) and second on medical data from hospital (progression CO2 in brain). We use these two data sets, because of their different properties. The "Mandarin" data demonstrates the periodic behavior (see Figure 1), whereas the "Brain" data set is strictly aperiodic (see Figure 5) from its nature.

The setup of the experiments can be found in the Section 4.1 and 5.1 of this paper. In sections 4 and 5 we are experimenting with short and medium term prediction and the final comparison and evaluation of results are given in Section 7.

2 The Analogues Complexing GMDH

Analogues Complexing is non-parametrical algorithm of GMDH. It is a multidimensional pattern search method that can be used for clustering, classifying, and predicting most fuzzy objects. For

prediction, for example, it self-selects several similar patterns relative to a given reference pattern and then uses their known continuations to form a prediction for the reference pattern.

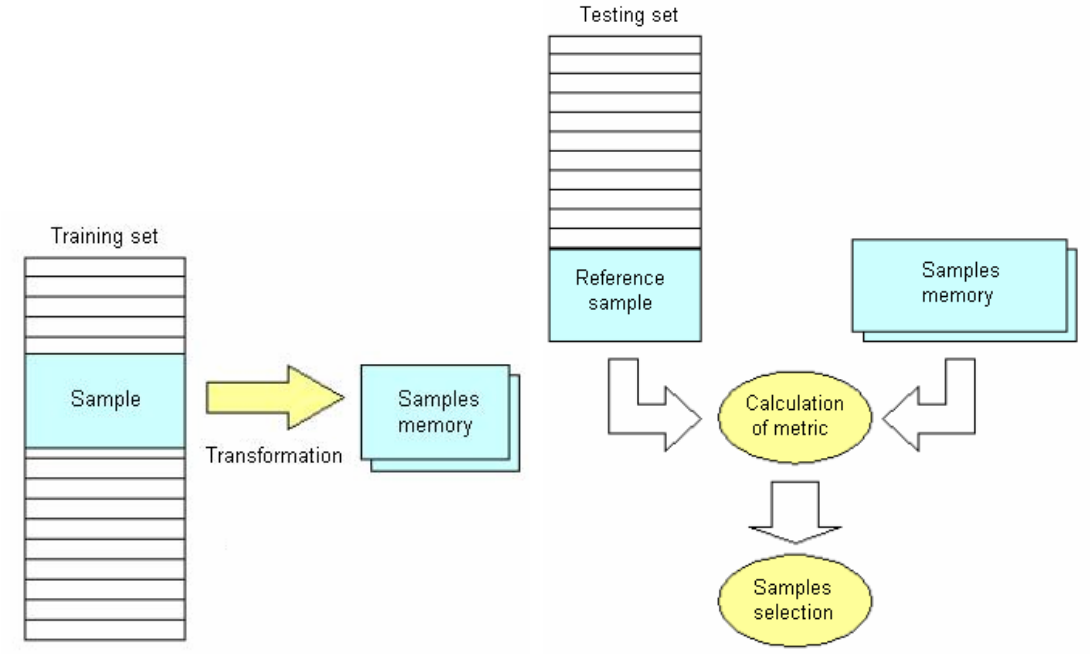

Fig. 1. The Analoques Complexing algorithm stores samples of historical signal (training set) in a memory. For the purpose of prediction, a reference sample is compared to samples in the memory. The continuations of selected sample(s) form the final prediction.

The AC method is frequently applied to short noisy data sets with fuzzy properties and aperiodic behavior. It is possible that for such data sets the AC method has better properties than parametric methods. In this paper we would like to find out, if the AC method is superior to parametric methods (e.g. GAME) when applied to time series with periodic and aperiodic properties.

3 The GAME method for time series prediction

The GAME states for the Group of Adaptive Model Evolution. This method is proceeds from GMDH theory specifically the Multilayer Iterative Algorithm (MIA) [2]. The GAME generates models inductively from a data set. The model grows from a minimal form during the learning phase, until the optimal complexity is reached. Starting from the first layer, a special genetic algorithm [7] evolves units in the layer. Units can be of several types – differing in the function transferring input signals to their output (linear, polynomial, sigmoid, etc.). The most successful units (fitness is computed using an external criterion – e.g. performance on the validation set) from the population of the genetic algorithm are frozen and form the first layer of the GAME model.

The method proceeds with next layers, evolving units which are most suitable for the given data set. The resulting model consists of heterogeneous units effectively interconnected in a feedforward network manner (see Figure 2).

The GAME method is primary designed for regression and classification tasks. The prediction is not so straightforward. A training data set has to be prepared from historical signal using the Sliding window approach [6]. This training data set is used to evolve GAME models capable of single value prediction (model has just single output). To predict more values, it is necessary to evolve more models –with the same inputs and different outputs.

The Figure 3 shows how the GAME model can be evolved and used for prediction of single future value (t+1).

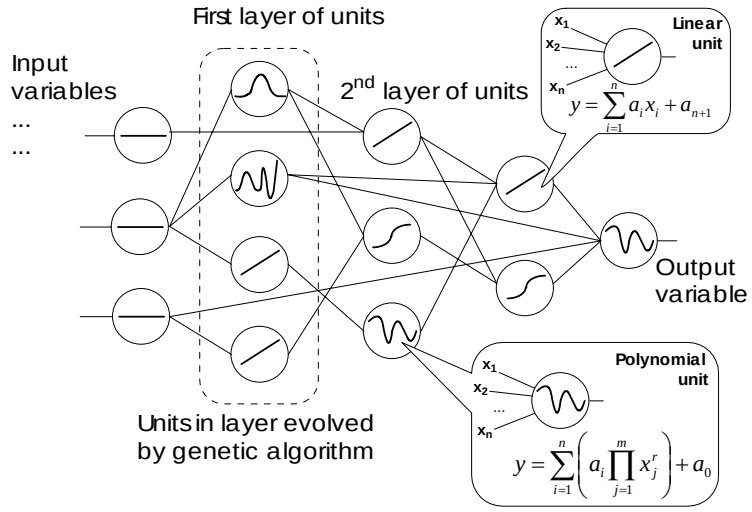

Fig. 2. An example of GAME model consisting of heterogeneous units interconnected into a feedforward network.

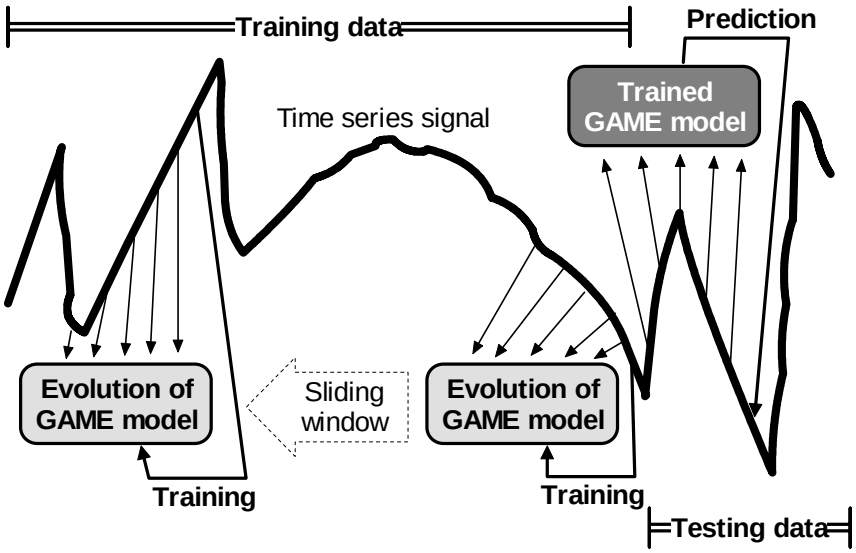

Fig.3. The scheme of time series prediction by means of GAME model. The model is first evolved using training data prepared from the time series using the sliding window approach, and then the prediction of the model is evaluated on testing data set.

4 Setup of experiments

We designed the experiments to find out whether the AC method predicts better than parametric models evolved by the GAME method.

The experiments were performed on two different time series data sets – one with periodic behavior and the second with aperiodic development of signal.

We used the implementation of the AC method implemented by Radek Pinc [1] and our open source environment FAKE GAME [8] allowing building GAME models.

At first, we used the same training sample for both methods and compared their prediction on a testing sample. Then we compared the performance of methods on several testing samples to get more accurate results.

5 Results on Mandarin data (periodic behavior)

The Mandarin data set (periodic time series) was provided by Hort Research Company, New Zealand. It contains measurements of mandarin tree water consumption. During a day, a tree consumes much more water than during the night (see Figure 4).

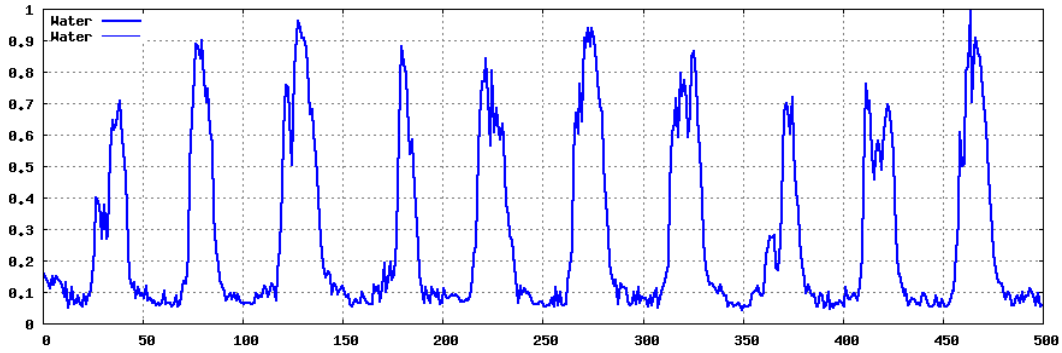

Fig. 4. Training set of mandarin data for prediction in both of methods

The training data set containing 500 measurements was used for both the GAME and the AC methods. The data set for testing contained 120 subsequent measurements. For medium and longer term prediction, we built several GAME models, each trained to predict the signal at certain time horizon.

The differences of model output and the target values were measured as the RMS error, which can be computed as:

$$\mathbf{RMS} = \sqrt{\frac{1}{n} \cdot \sum (y - d)^2}\ , \qquad (1)$$

where n is number of target values, y is real value from the testing set and d is a predicted value.

Tab.1. RMS error of prediction by GAME method on mandarin data, on the sample 51

		GAME		AC	
Index	Real data	prediction	$(y-d)^2$	prediction	$(y-d)^2$
51	0,2730	0,227599	0,0020612	0,124417	0,0220770
53	0,5340	0,595755	0,0038137	0,210808	0,1044528
55	0,7295	0,653557	0,0057673	0,457020	0,074245
57	0,9280	0,873399	0,0029813	0,617773	0,0962405
59	0,9260	0,861338	0,0041811	0,663593	0,0688572
61	0,8770	0,658491	0,0477461	0,775420	0,0103184
63	0,9160	0,860168	0,0031172	0,837687	0,0061329
65	0,7430	0,642948	0,0100103	0,832420	0,0079959
67	0,5580	0,470969	0,0075744	0,818920	0,0680792
69	0,3080	0,391431	0,0069606	0,644293	0,1130932
		RMS = 0,0970637		**RMS = 0,2390591**	

The Table 1 [1]explains, how the RMS error was computed for both GAME and AC methods and how these errors can be compared.

5.1 Setup of experiments

The GAME models had 30 historical values in their inputs (see Fig. 3). It means that the prediction by GAME is based on window of 30 reference samples in testing set and 20 future unknown samples which are predicted (for example: on the Table 1, there is a prediction of 20 future values from sample 51, and for inputs the reference samples 30-50 are used). The same at the Figure 5 – thick blue line 16-46 are the reference samples on the input, and outputs of models are samples 47-67 – red line. The thin blue line from sample 46 connects target values.

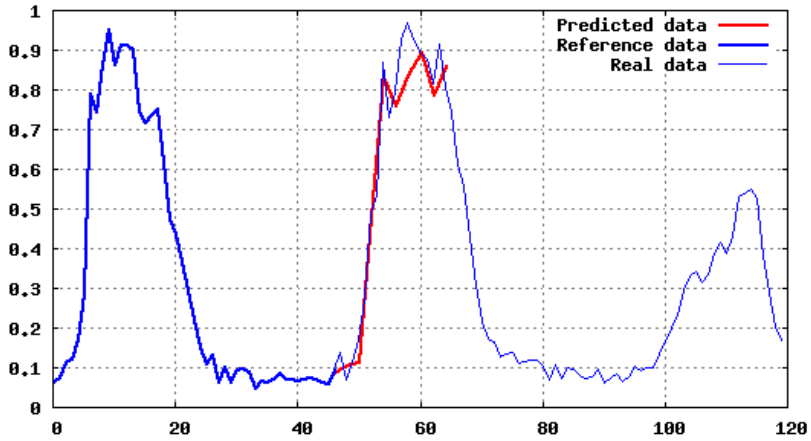

Fig. 5. Prediction of 20 future values with GAME on mandarin data on sample 46

In setup of GAME are used linear and polynomial neurons and perceptrons. As training method the Quasi Newton method was preselected.

The prediction by GMDH Analogues Complexing is based on the range 15 – 30 historical values. This method tries to predict the whole rest of the testing set (see Figure 6), but for the purpose of our experiment only 20 future values are used for comparison with the GAME method. In the setup of the GMDH AC method, we used the Normalization of samples and the Euclid metric for selection of the best samples from the training set (the most similar to the samples in reference window of the testing set).

The prediction displayed on the Figure 6 shows that the AC method has in some parts of the time series tendency to delay the real signal. This shift of signal significantly contributes to the RMS error and therefore the results of the AC method are much worse than that of the GAME method.

However it is clear that we cannot make a conclusion from one specific observation, therefore we performed additional experiment described in later sections to verify this preliminary result.

[1] There are only odd indexes in tables, because of constrained number of out values from GAME and effort to get longer time of prediction. The odd indexes in table of GMDH are there for better comparison with results of GAME.

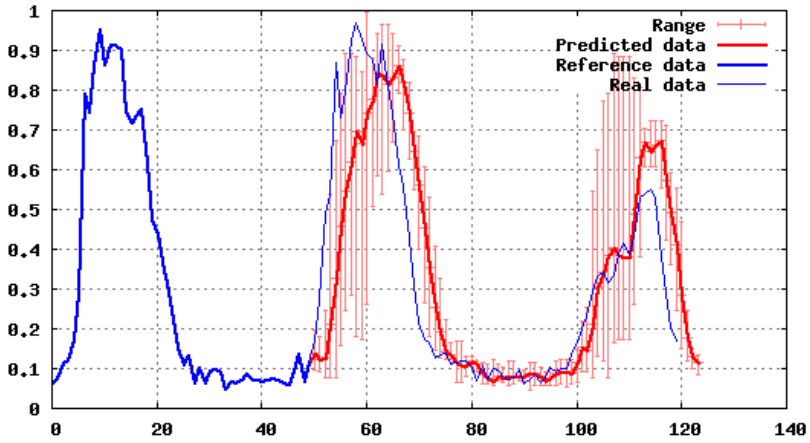

Fig. 6. Medium-term prediction with GMDH Analog Complexing on mandarin data.

6 Results on Brain injury data (aperiodic behaviour)

In this data set, we used the same methodology as for the Mandarin data set. The training sequence displayed on the Figure 7 clearly demonstrates the aperiodic properties of the time series.

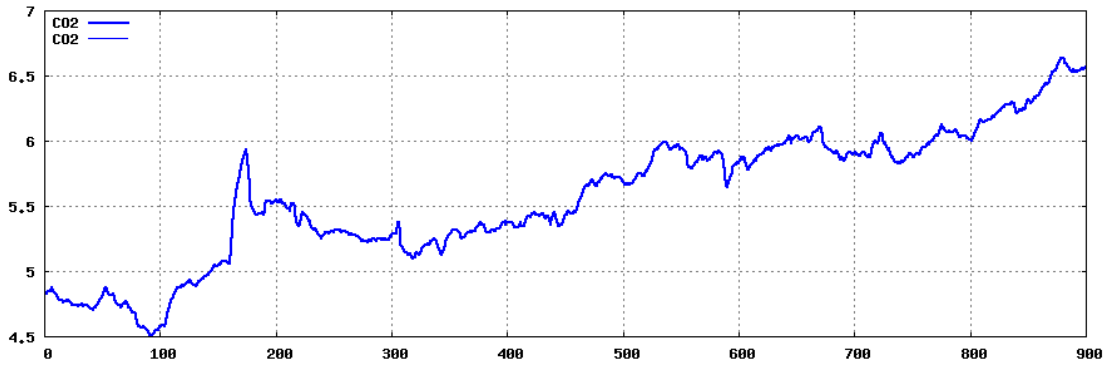

Fig. 7. Training set of medical data for prediction in both of methods

6.1 Setup of experiments

In the experiments on aperiodic data training set of 900 samples is used (Fig. 7). The setup of GAME and GMDH method is the same as the setup for the Mandarin data set in the previous section.

At first, we performed the same experiments as in the previous section – the prediction of 20 future values. The testing set consisted of 100 samples continuing from training set.

Results for the GAME and the AC method are displayed on Figures 8 and 9 respectively. The RMS error of the GAME model computed from prediction of 20 future values was 0,019049. For the GMDH AC method, the error was approximately twice as high (0,038889). Again, more confident results are given in the next section.

The Figure 10 demonstrates that the AC method is capable of medium and long term prediction. However the error of the prediction tends to increase with longer prediction horizon. We will verify this assumption in the next section.

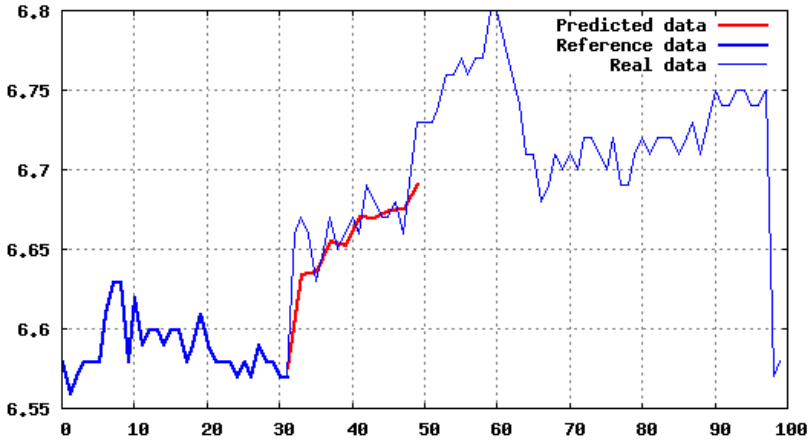

Fig. 8. Prediction of 20 future values with GAME on Brain Injury data, on sample 31

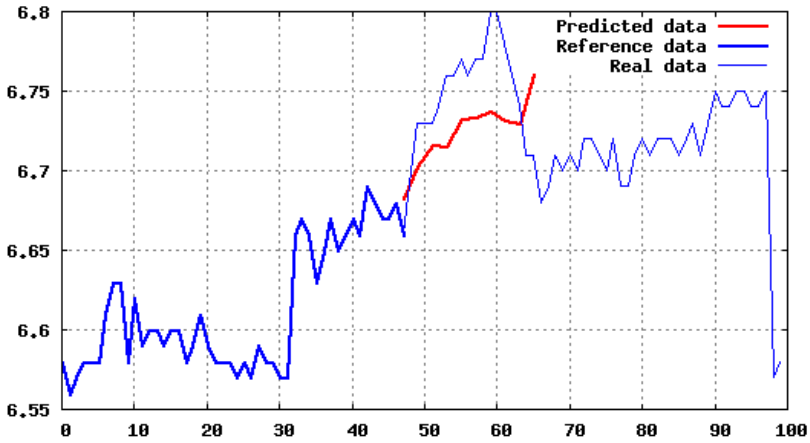

Fig. 9. Prediction with GMDH Analog Complexing on Brain Injury data, on sample 47

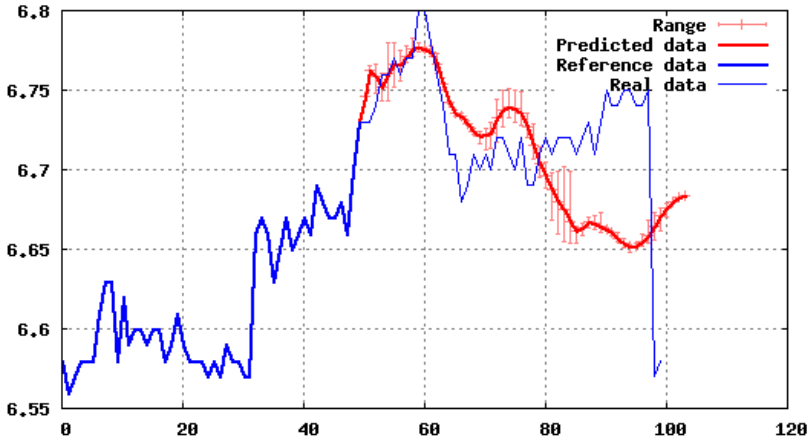

Fig. 10. Example of medium-term prediction with GMDH Analog Complexing

7 Evaluation of methods and results discussion

The Table 2 summaries average errors of prediction on the Mandarin and the Brain Injury testing data for several different testing sets. A sliding window method was used to generate 20 testing sets and reference sets from testing data. The same results are depicted on the Figure 11.

Tab.2. Comparison of RMS error of prediction 20 future values by GAME and GMDH (averaged from 20 measurements for different testing data sets)

RMS error	Mandarin data	Brain injury data
GAME	0,073994	0,038623
GMDH	0,158162	0,070403

The difference is evident. The prediction on periodic data using GAME method has better results than GMDH AC. The big RMS error of GMDH method is caused by translation predicted from real values along time axis, but the prediction using GMDH is easy for long term prediction.

The RMS errors on Mandarin data and the Brain injury data cannot be compared. Each data set has different properties and therefore RMS errors can be used just to compare the performance of individual methods on single data set.

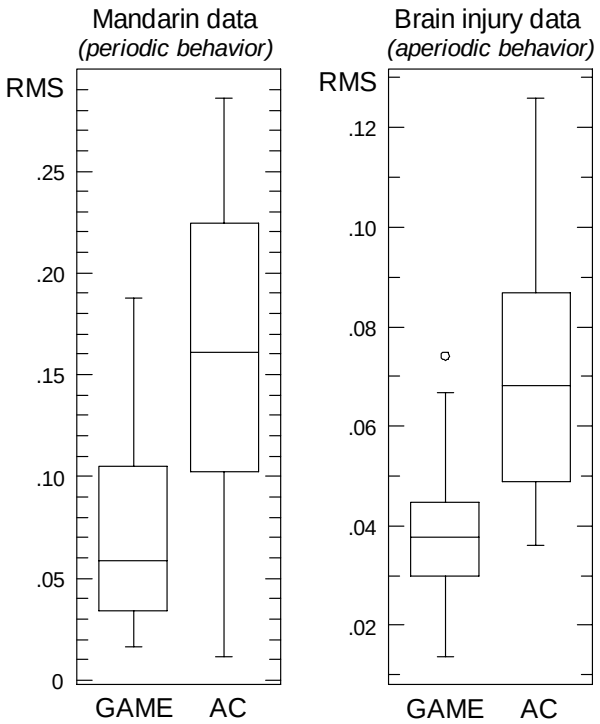

Fig. 11. The RMS errors' box plot: prediction by GAME and GMDH AC on Mandarin and Brain Injury data collected on 20 different testing sets.

We were interested, if there is a relationship between the error of the prediction and the time horizon (how many time steps to the future we are predicting). The Figure 12 shows that for the Analogues Complexing method, the error naturally increases with the distance of the target value in the future. Also the dispersion of errors increases. This behavior is apparent for both Mandarin and Brain injury data sets.

Very interesting are the results of the GAME method. For the Brain injury data, the error of the prediction stays on the same level even for the prediction of target value 10 time steps in future. Even more surprising is the decreasing trend of the errors dispersion. We will investigate this behavior in the future. For the Mandarin data set the behavior is similar except that the trend of errors dispersion is the opposite (that is more natural).

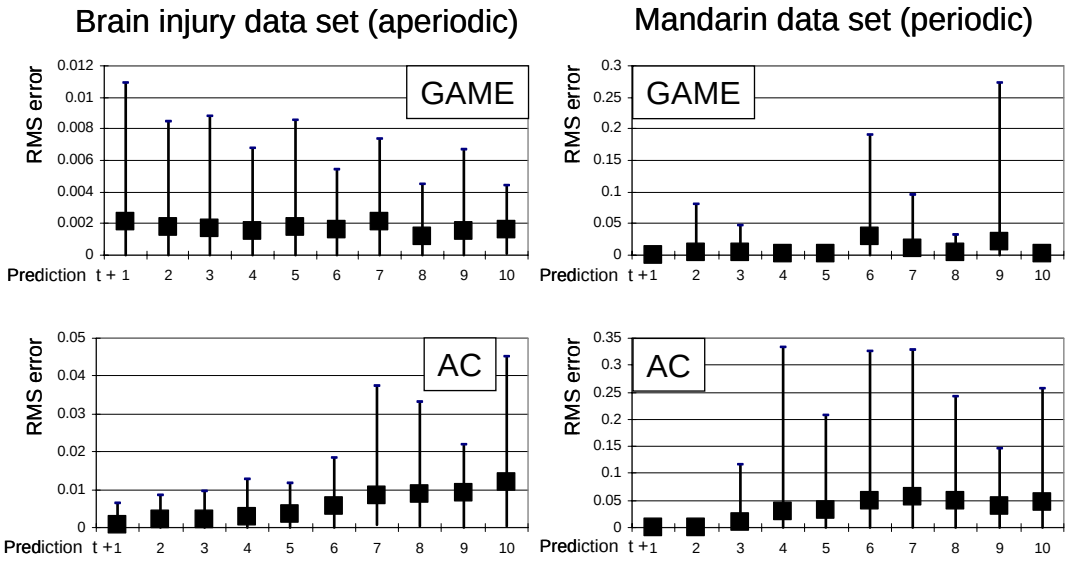

Fig. 12. Average, maximal and minimal RMS errors for 20 testing samples and different prediction intervals.

The Analogues Complexing GMDH method has similar behavior for both aperiodic and periodic data. The average and the dispersion of errors have growing trends with the time distance predicted to future.

8 Conclusion

In this paper we were looking for the answer to the question if parametric method (GAME) is better than non parametric method (GMDH AC) for the prediction of time series. We experimented with both periodic and aperiodic time series.

Our results showed that the GAME models are superior to the AC method for both periodic and aperiodic time series (GAME is twice as accurate as GMDH AC).

For the short-term and medium-term prediction the GAME is more accurate than the GMDH AC. The questionable remains the prediction for longer time, which has not been tested yet. The performance of the GAME method is quite promising for longer time horizon, because the error of this method is not increasing much with the distance of the prediction (see Figure 12). This will be subject of our further research.

9 Acknowledgements

We would like to thank to Dr. Richard Brzezny from Dept. of Neurosurgery, The Motol University Hospital in Prague, Czech Republic for the Brain Injury data set and to Dr. Phil Prendergast from Hort

Research company, Kerikeri, New Zealand for the Mandarin data set. Thanks to Radek Pinc for his implementation of the Analoques Complexing GMDH method.

This research is partially supported by the internal grant of the Czech Technical University in Prague (CTU0715313), by the grant Automated Knowledge Extraction (KJB201210701) of the Grant Agency of the Academy of Science of the Czech Republic and the research program "Transdisciplinary Research in the Area of Biomedical Engineering II" (MSM6840770012) sponsored by the Ministry of Education, Youth and Sports of the Czech Republic.

References

[1] Pinc R.: Diploma thesis: Implementation of GMDH Analog Complexing, 2005 (in Czech language)
[2] The short description of the Analoques Complexing Algorithm (AC GMDH) online at http://www.gmdh.net/GMDH_ana.htm
[3] The short description of the Multi-layered Iterative Algorithm (MIA GMDH) online at http://www.gmdh.net/GMDH_mia.htm
[4] Ivakhnenko,A.G. An Inductive Sorting Method for the Forecasting of Multidimensional Random Processes and Events with the Help of Analogues Forecast Complexing. *Pattern Recognition and Image Analysis*, 1991, vol.1, no.1, pp.99-108.
[5] Kordík P.: GAME - Group of Adaptive Models Evolution, dissertation thesis proposal DCSE-DTP-2005-07, FEE, CTU Prague, 2005
[6] Chu, C.J.: Times series segmentation: a sliding window approach, Information Sciences—Informatics and Computer Science: An International Journal archive 85, p.147-173, 1995.
[7] Mahfoud, S.W.: Niching Methods for Genetic Algorithms (95001), Technical report, Illinois Genetic Algorithms Laboratory (IlliGaL), University of Ilinios at Urbana-Champaign, 1995
[8] The FAKE GAME environment for the automatic knowledge extraction, available online at: http://neuron.felk.cvut.cz/game