Model of Prediction Error in Chaotic and Web Driven Business Environment

Franjo Jović*, Alan Jović **
** Faculty of Electrical Engineering, University of Osijek, franjo.jovic@etfos.hr*
*** Faculty of Electrical Engineering and Computing, University of Zagreb, alan.jovic@fer.hr*

Abstract. *Prediction of time series data. Comparison between chaotic and web driven business transactions. Technology of artificial neural networks, genetic algorithms and Quacol algebra. Takens equation for sampling, width and prediction horizon. Prediction technique: functional or stochastic model. Fit composition of prediction functions. Predicting data from chaotic behavior in a ferroresonant circuit. Prediction of web driven trading process. Prediction error: sampling interval, correlation, time horizon. Modeling prediction error in Quacol algebra using triangle inequality. Discussion of practical results.*

Keywords. explicit model, qualitative algebra, prediction model, rank continuity principle

1. Introduction

Prediction is usually treated through Bayes formula. Although probability of an outcome relies on past data it is bravely hypothesized that the continuation of the past can be predicted by modeling. In most crucial cases where prediction is pragmatically sought this condition is not fulfilled. Two schools of prediction philosophy are usually followed: using stochastic or functional patterns of previous data behavior. There are odds in favor of each one of them. We will make a model of the data behavior using different techniques and thus empirically try to predict future data.

This work consists of: comparison of linear and nonlinear models, case of chaotic data prediction, comparison between deterministic chaos and business transactions, introduction of prediction technologies: artificial neural networks, genetic algorithms and circular qualitative correlation algebra (Quacol algebra). Prediction by a fit composition of analytical functions. Practical results of prediction of data from a chaotic ferroresonant circuit and from a web driven trading process. Analyses of the prediction error from the standpoint of sampling interval, correlation and prediction time horizon. Modeling prediction error in Quacol algebra using triangle inequality.

2. Prediction modeling technologies

2.1 Linear and nonlinear models

The most simple prediction model is a linear model. Here the future x_k component of the signal is given with the expression

$$x_k = \sum_{1}^{i} \alpha_i x_{k-i} + w_k \qquad (1),$$

where:

x_{k-i} - signal component determined in the *i*-th previous prediction interval

w_k - unknown white noise component at the prediction instant

α_i - coefficients of the time series expansion of previous time instants.

This model is still in use for various technical estimations and prediction purposes such as given in Smith J.A. US Patent 5,606,505 and Wojsznis W.K. US Patent 5,568,378. Different approach is presented by Frederic D. Fox and coauthors who considered a prediction method using regression analysis and artificial neural network [1]. Using long term data the system predicts weather from three days to 15 months in advance with typical accuracy of weekly weather forecasts around 70%. Using correlation of previous weather data and POS store transactions data the system advises retailer on the managerial actions to be taken. In such a way hidden patterns of weather behavior have been pre-selected by the ANN.

In order to generate patterns in advance the method has been proposed by John Koza [2] whereby a composition of problem solving entities has been generated and combined in a genetic algorithm version of the problem solution. Such a combination of function can be

used for training of the prediction possibilities which was not developed by the above mentioned author. Still genetic algorithms can be used for constructing models fit for prediction.

Models of chaotic processes are the most difficult for prediction. Takens has proposed the equation for sampling width, prediction horizon, and functional for a chaotic process, but without stating neither the horizon accuracy nor the functional nature [3]. Perlovsky on the other side advocates functional approach to modeling unknown processes in the nature and human activities [4].

2.2 Quacol (qualitative correlation) algebra predictor

2.2.1 Qualitative explicit model

Qualitative data can be obtained from quantitative data by a simple ranking procedure. The positive ranking assignment is applied to set of variable data such as measurement data obtained in successive time intervals, i.e. measurement vector $v_1 = (3.69, 7.15, 4.37, 15.73, 0.18)$ is transformed into its corresponding n-point graph V_1 = (4, 2, 3, 5, 1), v_2 to V_2 etc. Any desirable function that is investigated can be defined as goal function, g_1 = (10.06, 27.97, 15.28, 37.66, 0.12), and it is transformed into the corresponding goal n-point graph G_1 = (2, 4, 3, 5, 1). Their rank correlation equals to

$$\rho_{V,G} = 1 - \frac{6\sum \Delta^2}{n(n^2-1)} \qquad (2),$$

where $\sum \Delta^2$ equals the sum of correspondent squares of rank differences for the n-point graph. Thus for the illustrated series $\rho_{V_1,G} = 0.56$.

The manipulation with greater number of variables and their inverses results in a variable with the highest rank correlation coefficient. Difference in ranks between this variable and goal function is used to give another artificial goal variable to enter as the algebraic counterpart of the missing rank difference, i.e. there is a rank difference between goal function and model variable that equals to

$$\Delta(G_1 - V_1) = (-2, 2, 0, 0, 0) = G_{2improper} \qquad (3),$$

where the subscript "improper" designates rank difference function, i.e. the value that has not been yet properly ranked. After shifting (3) by adding a positive constant vector such as 2 the corresponding quantitative function $g_2 = (0, 4, 2, 2, 2)$ can be obtained. Mixing g_2 values with a positively defined strictly increasing additive „background noise" $V_n = (0.01, 0.02, 0.03, 0.04, 0.05)$ and after ranking one obtains difference goal function $G_{2n} = (1, 5, 2, 3, 4)$.

After that the following relation holds [5]:

$$G_1 \; corresponds \; R(v_1 + k \; (corresponds \; R(G_{2n})) \qquad (4),$$

where the *corresponds* operator searches the most appropriate variable according to its ranks to the corresponding G_1 or G_{2n} goal function ranks, $R(.)$ is the rank operator. Let us assume that the variable V_2 maximally corresponds to the G_{2n} goal function according to relation (1). Then there follows:

$$G_1 \; corresponds \; R(v_1 + kv_2) \qquad (5).$$

The algebraic relation in (5) contains the constant k for which can be proven that the relation in parenthesis can have the maximum rank correspondence being a convex function on (2), thus

$$k = \max(G_1 \; corresponds \; R(v_1 + kv_2)) \quad (6).$$

The implicit value for k has to be calculated for each combination of variables in the whole variable data set.

2.2.2 Quacol algebra

Two principles of the Quacol modeling algebra, such as given for example in equation (5), have to be defined:

First is the principle which states that any n-point graph should not have any equal ranks, e.g. $V_k(1, 2, 3.5, 3.5, 5)$ is not allowed.

The second is the principle of continuity of the n-point graph for specific algebraic operations of multiplication and division.

Definition 1. The rank value of two values in any n-point graph or goal function are not allowed to be equal. Single valuedness of the ranks in any variable is a practical demand that avoids unnecessary calculation of the complex correlation formula for equal ranks. The possible equal data in any variable are solved by the addition of a very small amount of noise to each data in each variable, and theoretically to each varable pair algebra as well.

Definition 2. Any algebraic operation between any two variables can not influence on the rank continuity of any variable. This is a fundamental demand that changes the multiplication and division operation in Quacol algebra where the operations are defined according to Table 1. The proof of result from Table 1 is fairly simple: it stems from a theorem in Quacol algebra that states:

Table 1. Multiplication and division operations in Quacol algebra

Result of multiplication (v1*v2) and division (v1/v2) operation	v1 data sign positive	v1 data sign negative
v2 data sign positive	Positive	Negative
v2 data sign negative	Negative	Negative

Theorem 1. Any n-point graph G is invariant to the linear scaling operation with vector c, $c_i \in R, c_i > 0$

$$c\ op\ G \equiv G \qquad (7),$$

where $op = \{+,-,*,/\}$. For example, taking a variable $v_3 = (2.5,4,5,1); V_3 = (2,3,4,1)$ and subtracting a constant vector $c = [2, 2, 2, 2]$ one obtains $v_3 - c = (0.5,2,3,-1)$; $R(-c + v_3) = R(v_3)$ according to Theorem 1.

2.2.3 Quacol predictor

Let us define as the prediction goal function any desirable goal function g_k of the depth n, where k is the total number of variables of a system, including the goal function, i.e. $\{g_k, v_i\}, i = 1,\ldots,k-1$. Goal model can then be expressed as:

$$M_{g_k} = \{op\}_m \left[v_{m,i}^{\{ord,inv\}}, v_{m,j}^{\{ord,inv\}}, op_m, k_m \right]_{mean(m)} \qquad (8),$$

where $\{op\}_m$ is a sequence of m algebraic operations performed on model in square brackets; $v_{m,i}^{\{ord,inv\}}$ is the first variable in the m-th model, with index i, $i = 1,\ldots,k-1$, that may or may not have inverted values; $v_{m,j}^{\{ord,inv\}}$ is the second variable in the m-th model, with index j, $j = 1,\ldots,k-1$; op_m is the algebraic operation between i-th and j-th variable and k_m is the weight of the second variable. All of the operations are performed on the variables that have been normalized to a common mean value for that model, denoted by $mean(m)$.

An example of the model is: $M_{g4} = \dfrac{[v_1 + 0.5inv(v_2)]_{mean1}}{[v_2 - 0.25inv(v_3)]_{mean2}}$.

By using a prediction vector x_{n+1} we can predict a future value of g_k when x_{n+1} is added as a last component to each $v_i, i = 1,\ldots,k-1$; thus $predictor(g_k) = M_{g_k, x_{n+1}}$.

Number of iterations following the procedure described under expression (8) is sometimes limited due to numeric instability of the procedure [6].

3. Practical investigations

Predictor limits were tested under following constraints:

i) There was only one system variable and that one has to be predicted from past values. This is illustrated for the voltage signal of the ferroresonant circuit.

ii) The time horizon was tested for small variable set $k = 4$, the case of trading variable prediction

iii) The prediction precision was tested for short and long prediction interval d for trading variables, $k = 4$.

3.1 Predicting chaotic behavior of the ferroresonant circuit

Synthetic functions have been used such as $v_{k-1} * v_{k-2}$ or $\sqrt{v_{k-1}}$ or similar analytical forms.

Prediction data for ranks of the ferroresonant circuit are given in Table 2. Mean prediction error of the linear model was around 277% and prediction error of the Quacol synthesized predictor was around 108%. The actual voltage levels were between -0,1969V and 0,3464V.

Table 2. Prediction ranks and rank ranges for the ferroresonant circuit in chaotic behavior [7]

	T1	T2	T3	T4	T5	T6	T7	T8	T9	T10	T11	T12	T13	T14	T15	T16	T17	T18
Goal rank	1	2	5	7	8	9	10	11	12	13	14	16	17	18	15	6	4	3
Quacol model rank	2	1	5	8	6	9	10	11	12	13	14	16	17	18	15	7	3	4
Linear model rank	3	2	1	16	6	9	12	7	10	11	13	14	15	17	18	5	1	4
Quacol model rank range	1-3	1-3	4-6	6-9	5-9	8-10	9-11	10-12	11-13	12-14	13-15	15-17	16-18	17->17	14-16	5-8	2-5	3-5

The worst case for linear predictor was predicting 2.5929V instead of 0.00825V and the worst case of the Quacol predictor was predicting the value between 0.024V and -0.0176V instead of 0.01892V.

3.2 Determining the prediction horizon for small number of trading variables

Four trading variables were observed: closing, opening, high and low values. The intervals of scanning were one hour and the time duration was 24 hours. The 25-th value was predicted with different accuracy, Table 3, for ten different trading situations (last three or four digits were given). Synthetic analytical variables were not used.

Table 3. Trading forecasting for ten cases for „opening“ variable

Case	1	2	3	4	5	6	7	8	9	10
Predicted value	<2105	285-325	585-605	>945	225	>205	385-415	605-615	885-905	385-415
Prediction class	B	A	A	B	C	A	C	A	A	B
Actual value	1995	325	595	985	235	225	375	605	895	435

Prediction class is formed according to rank correlation coefficient span: A>0.99, B (0.97-0.99) and C (0.95-0.97). When the horizon was extended to two-hour periods the correlations have decreased to the values between 0.70 and 0.80 (prediction class F) or smaller thus decreasing the accuracy of the prediction for more than three ranks.

3.3 Determining prediction accuracy for small number of prediction variables and different variable lengths

Two cases have been studied: accuracy of *n*=25 data series and *n*=80 data series. Data on shorter model showed overall accuracy around 20%. Prediction for an 80 data series model are given in Table 4. Synthetic analytical variables were not used.

4. Prediction error modeling

Goal G_1 and goal difference functions G_{2n} are linear independent variables, meaning that they are principally collected from mutually inverse variables and calculated in geometric way toward goal function fulfillment.

Definition 3. An error in prediction $Error(m)$ using Quacol algebra is consistent if for each model member of the n-point graph (v_1) and for every successor variable n'-point graph

($v_2 = v_1^{-1}$) obtained from its difference toward the goal function *g*, the estimated error of reaching the goal from n-point graph is no greater than the error of obtaining the goal from getting to n' plus the estimated error of reaching the goal from n':

Table 4. 80 days prediction data for small number of prediction variables (Croatian stock market)

Variable	Prediction	Real value	Variable span	Relative error	Correlation coefficient	Prediction class
High	912	913	880-929	-2%	0.78	F
Low	901-910	901	870-920	+18%	0.79	F
Opening	902-920	910	872-925	-15% +19%	0,76	F
Closing	919-925	910	880-924	20%/34%	0,72	F

$$Error(m) \le Error(v_1) + Error(v_2) \quad (9).$$

Thus simply calculating the model from variable and its inverse gives

$$R(v_1 + v_1^{-1}) = R(3.96, 7.29, 4.60, 15.79, 5.74) = \\ = (1, 4, 2, 5, 3) \quad (10),$$

showing a better correlation of the goal and model (10) with $\rho = 0.7$ as compared with previously obtained $\rho = 0.56$ for v_1 only.

5. Discussion

Prediction accuracy and correlation of models are highly connected. This stems from ideal prediction case where the model behavior of the goal function has been completely discovered ($\rho = 1$). Prediction error is then defined with the interval closest to the rank of the expected value. Ideally for 100 equidistant values of the goal function and completely discovered goal function the accuracy is +/- 1% . Realistical expectations are far less favorable. Neither there are long enough data series that are without large chaotic behavior nor there is any linearity in the goal data distribution. Still predictions of smaller data series (>25 data series) can be expected with about 10% accuracy which can be favorable for many practical applications.

6. References

[1] Fox F.D. et al., "System and Method for the Advanced Prediction of Weather Impact on Managereal Planning Applications", US Patent 5,521,813 , May 28, 1996.

[2] Koza J.R., "Non-linear Genetic Algorithms for Solving Problems by Finding a Fit Composition of Functions", PCT/US91/01970 , March 25, 1991.

[3] Diambra L., Plastino A., "Modelling Time Series Using Information Theory", *Physical Letters A* 216, pp. 278-282, 1996

[4] Pavlovsky L.I., "Toward Physics of the Mind: Concepts, Emotions, Consciousness, and Symbols", *Physics of Life Reviews* 3, pp. 23-55, 2006

[5] Jović, F., "A Circular Qualitative Algebra", *Casys*, 8:213-225, 2001

[6] Jović, F., Jović, A., Rajković, V., "A Contribution to the Investigation of Algebraic Model Structures in Qualitative Space", CTS&CIS pp. 49-53, Mipro 2006

[7] Pelin, D., Fischer, D., Flegar, I., „Observing Chaos in a Ferroresonant Series Circuit", International Conference on Electrical and Electronics Engineering, Turkey, December, 1999