Modeling Prediction using Quacol Algebra in Web-driven Business Environment

Jović Franjo; Jović Alan

Abstract— ***Prediction of time series data for chaotic and web driven business transactions. Prediction technologies: linear regression, artificial neural networks, genetic algorithms and Quacol algebra. Dilemma in prediction technique: functional or stochastic model. Fit composition of prediction functions – Quacol predictor model. Elaboration on rank exclusivity, on continuity of n-point graph and theorem on sign independent algebraic operations. Practical prediction data from chaotic behavior in a ferroresonant circuit. Elaboration on prediction of web driven trading process. Modeling prediction error in Quacol algebra using triangle inequality.***

Index Terms— ***explicit model, qualitative algebra, prediction model, rank exclusivity, rank continuity***

I. Introduction

PREDICTION is usually treated through probabilistic Bayes formula. If we know the probability of outcome B and the joined probability of occurrence of both the outcomes A and B, we can calculate the conditional probability that outcome B occurs if outcome A has occurred [1]. That is a simple and powerful set-based prediction method. Although probability of such 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: those using stochastic or those using 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 based on algebraic construction of various data-based analytical prediction forms.

Manuscript received January 30, 2007.
F. Jovic is with the Faculty of Electrical Engineering, University of Osijek, Hrvatska (e-mail: franjo.jovic@etfos.hr).
A. Jovic is with the Faculty of Electrical Engineering and Computing, University of Zagreb, Hrvatska (e-mail: alan.jovic@fer.hr)

This work consists of: comparison of linear and nonlinear models, case of chaotic data prediction, comparison between deterministic chaos and business transactions, introduction to prediction technologies: artificial neural networks, genetic algorithms and circular qualitative correlation algebra (Quacol algebra) [2]. Prediction by a fit composition of analytical functions in Quacol algebra will be given in more detail as the focus of this work. Results are also given by 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 are discussed. Also, a pragmatic modeling method for prediction error in Quacol algebra and the consistency of error is defined.

The organization of this work is as follows: Chapter 2 deals with different prediction methods and chapter 3 presents the results of prediction in several applications. Chapter 4 shows the approach for estimating prediction error in Quacol algebra.

1. Prediction Modeling Technologies

2.1 Linear and Nonlinear Models

The most simple prediction model is a linear autoregression 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 (assumed normal distribution)

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

Standard error of estimation, used in prediction, is given by:

$$s = \sqrt{\frac{\sum (x_i - x_i')^2}{n-k-1}} \quad (2)$$

where n is the number of samples, k is the number of preceding values of observed value x_i, x_i' are the theoretically predicted values.

This model is still in use for various technical estimations and prediction purposes such as given in [3] and [4]. Different approach is presented by Fox and coauthors who considered a prediction method using regression analysis and artificial neural network [5]. 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 Koza [6] 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 functions 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 because of their dynamic nonlinearity. Diambra [7] 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. Perlovsky [8] on the other hand advocates functional approach to modeling unknown processes in the nature and human activities.

2.2 Quacol (Qualitative Correlation) Algebra Predictor

2.2.1. Qualitative Explicit Model

Qualitative data in Quacol algebra approach can be obtained from quantitative data by a simple ranking procedure. The positive ranking assignment is applied to a set of variables. When ranked, these variables are called n-point graphs (or n-graphs) in Quacol algebra. The ranking is usually performed on a set of time series data, however, the ranking can be applied to any quantitative variable. For example, a 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, 1, 5), a v_2 to V_2 etc. Any desirable variable that is investigated can be defined as goal function, e.g. g_1 = (27.97, 10.06, 15.28, 37.66, 0.12) is transformed into the corresponding goal n-point graph G_1 = (2, 4, 3, 1, 5). Spearman rank correlation coefficient for ordinal variable equals to [9]:

$$\rho_{V,G} = 1 - \frac{6\sum \Delta^2}{n(n^2-1)}, \quad (3)$$

where $\sum \Delta^2$ equals the sum of correspondent squares of rank differences for two n-point graphs, Thus for the illustrated series $\rho_{V_1,G_1} = 0.6$.

The selection performed according to (3) from a greater number of variables and their inverses results in an n-point graph with the highest rank correlation coefficient. Difference in ranks between this n-point graph and the goal function n-point graph is used to generate another artificial goal function to be entered as the algebraic counterpart of the missing rank difference, i.e. this is a rank difference between goal function and model variable, equal to

$$\Delta(G_1 - V_1) = (-2, 2, 0, 0, 0) = g_{2improper}, \quad (4)$$

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

After that the following relations hold, adapted from [10]:

$$G_1 \; corresponds \; R(v_1 + kv_i),$$
$$v_i \; corresponds \; R(G_{2n}), \quad (5)$$

where R(.) is the already explained rank operator and the *corresponds* operator first searches the most similar variable v_i according to its ranks to the corresponding G_{2n} goal

function ranks. After v_i has been found according to maximum of rank correlation coefficient amount (3), then a search for k is performed such that it minimizes the difference between the ranks of the sum $(v_1 + kv_i)$ and G_1.

2.2.2. Quacol Algebra

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

First is the principle of rank exclusivity 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. Rank correlation coefficient for equally ranked values would have to be calculated using an adapted formula [9,ibid], which is generally not used because of somewhat higher calculation demand for longer data series .

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

Definition 1. (Rank exclusivity)

The rank values of two values in any n-point graph or goal function are not allowed to be equal, i.e. $R(v_i(j)) \neq R(v_i(k)), \forall i, j \neq k$.

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 variable pair. The addition of noise enforces the distinction of values, with the price being the decrease in determinism of models with highly similar variable values and the gain is the ability to rank the values more efficiently.

Definition 2. (Continuity of n-point graph)

Any algebraic operation between any two variables can not influence on the rank continuity of any particular 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:

Theorem 1. (Sign independent algebraic operations) If the multiplication and division operations in Quacol algebra are defined according to Table 1, then the ranking operation performed on variables in positive domain is the same as one performed on variables with no restrictions on the domain, i.e.,

$$R(v_1 \, op \, v_2) = R((v_1 + c_1) \, op \, (v_2 + c_2)),$$

$$\text{if } (v_1 + c_1)_i > 0 \text{ and}$$

$$(v_2 + c_2)_i > 0, \; c_{xi} > 0, c_{xi} = c_{xj}, \forall i, j, \quad (6)$$

where $op = \{+, -, *, /\}$ is executed upon vector components and c_x are constant vectors. For example, let us take two variables: $v_1 = (2.5, -4, -5, 1); V_1 = (1, 3, 4, 2)$ and $v_2 = (-3, -4, 2, 3); V_2 = (3, 4, 2, 1)$.

If we perform multiplication $v_1 v_2$ according to Table 1 and rank the result, we obtain: $R(v_1 v_2) = R(-7.5, -16, -10, 3) = (2, 4, 3, 1)$.

If we apply the traditional definition of multiplication operation, we would obtain:

$R'(v_1 v_2) = R'(-7.5, 16, -10, 3) = (3, 1, 4, 2)$.

We perform a “lifting” operation upon variables v_1 and v_2 such that we translate their values into

TABLE 1: MULTIPLICATION AND DIVISION OPERATIONS IN QUACOL ALGEBRA

$v_3 = v_1 v_2$ or $v_3 = \frac{v_1}{v_2}$	$v_{1i} \geq 0$	$v_{1i} < 0$
$v_{2i} \geq 0$	$v_{3i} \geq 0$	$v_{3i} < 0$
$v_{2i} < 0$	$v_{3i} < 0$	$v_{3i} < 0$

the positive domain, e.g. $v_1 + (6,6,6,6) = (8.5,2,1,7)$ and $v_2 + (5,5,5,5) = (2,1,7,8)$. We than perform the same multiplication operation upon these altered variables and rank them. Thus, we obtain:

$(v_1 + c_1)(v_2 + c_2) = (17,2,7,56)$ and

$R(17,2,7,56) = (2,4,3,1)$.

This is the same result as without raising the variables' values into the positive domain.

It should be again noted that it is irrelevant whether we perform the alteration of the variables and calculate them in positive domain or we use the special rules for multiplication and division according to Table 1, because the end result is the same. Caution has to be exerted on the values of the variables if standard operations are used, because their values would have to be positive in that case.

2.2.3. Quacol Predictor Model

Let us define as the prediction goal function any desirable goal function (variable) 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:

$$M_{g_k} = \{op\}_m \left[v_{m,i}^{\{ord,inv\}}, v_{m,j}^{\{ord,inv\}}, op_m, k_m \right]_{mean(m)} \quad (7)$$

where $\{op\}_m$ is a sequence of m algebraic operations performed on model in square brackets, with respect to theorem 1; $v_{m,i}^{\{ord,inv\}}$ 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: $M_{g4} = \dfrac{[v_1 + 0.5inv(v_2)]_{mean1}}{[v_2 - 0.25inv(v_3)]_{mean2}}$.

In circular Quacol algebra, M_{g4} could be declared as a new variable and entered into the next cycle of goal estimation [10].

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 $v_i, i = 1,\ldots,k-1$ (all of the variables except g_k), thus $predictor(g_k) = M_{g_k, x_{n+1}}$.

Number of iterations following the procedure described under expression (7) is sometimes limited due to numeric instability of the procedure because of repetitious increase of the differences and mean values in the algebra [11].

2. PRACTICAL INVESTIGATIONS

Predictor limits were tested under following unfavorable conditions:

In chapter 3.1, there was only one system variable and that one had to be predicted from its past values. This is illustrated for the voltage signal of the ferroresonant circuit [12].

In chapter 3.2, the time horizon was tested for small variable set, $k = 4$, the case of trading variable prediction.

In chapter 3.3, the prediction precision was tested for different prediction interval d for small trading variable set, $k = 4$.

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, by a win – lose method.

Prediction data for ranks of the ferroresonant circuit are given in Table 2. Mean prediction error of the linear model was around 277% and the mean prediction error of the Quacol synthesized predictor was around 108%. The actual voltage levels were between -0,1969V and 0,3464V. 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 -0.0176 and 0.024V instead of 0.01892V.

3.2. Determining the Prediction Horizon For Small Number of Trading Variables

Four trading variables were observed: opening,

TABLE 2: PREDICTION RANKS AND RANK RANGES FOR THE FERRORESONANT CIRCUIT IN CHAOTIC BEHAVIOR [10]

	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

TABLE 3: TRADING FORECAST FOR TEN CASES OF THE „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

closing, high and low values of a stock market index. The intervals of scanning were one hour and the time duration was 24 hours. The 25-th value was predicted with varying accuracy, Table 3, for ten different trading situations (last three or four digits were given). Synthetic analytical variables were not used, because multivariable case is less sensitive to such improvements. 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 prediction accuracy.

3.3. 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=75 data series. Data on shorter model showed overall accuracy around 20% and are not considered here. We present the results for the longer data series. A resulting model n-point graph for a 75 days period is given in Figure 1. Data on prediction accuracy are given in Table 4. When we predicted the value for two days horizon, we took every second day in consideration, and for three days, every third day was taken. Table 4 shows that when one increases the prediction horizon, the error also shows a geometric increase. Obviously, it is more difficult to predict the value of a stock for more than one day in advance.

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.

A relative error of predictive model m in Quacol algebra is calculated using standard formula:

$$Error(m) = \frac{\bar{x}_{calc} - x_{real}}{x_{real}}, \quad (8)$$

where $\bar{x}_{calc}$ is the average value of the rank interval for the goal variable, i.e. "Predicted value" in Table 3, x_{real} is the real value of the

goal variable.

Finally, we define the consistency of an error of prediction in Quacol algebra.

Definition 3. (Error consistency)

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' (triangle inequality):

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

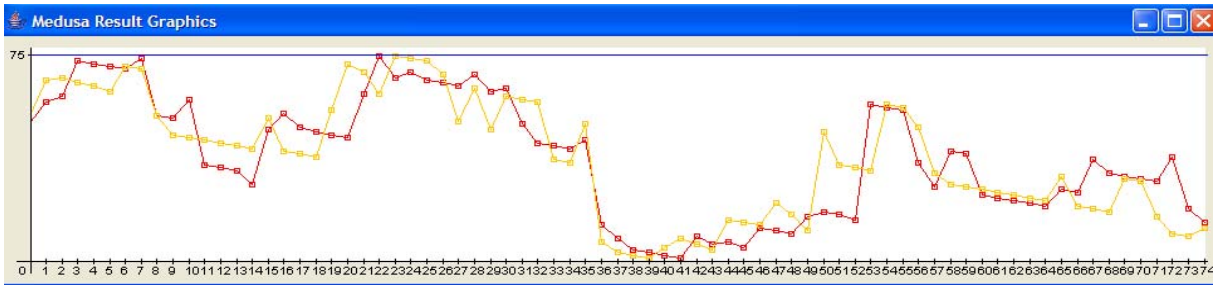

Figure 1. Model rank graph for 75 days period, yellow is the goal variable (opening), red is the model. MODEL(opening) = high + 0.90low

TABLE 4. 75, 38 AND 25 PERIODS PREDICTION DATA FOR ATPL STOCK ON THE CROATIAN STOCK MARKET*

Variable	Prediction	Real value	Variable span	Relative error	Correlation coefficient	Prediction horizon
Opening (76)	906-910	910	872-925	-3,78%	0.88	1 day
Opening (39)	906-906	909,99	872-925	-7,53 %	0.84	2 days
Opening (26)	901-901	909,99	872-925	-16,96 %	0,88	3 days

* Data for spring 2007 period;

5. DISCUSSION

Prediction accuracy and correlation of models are highly connected. It can be observed from (9) that prediction error is lower in simpler models, although the error is not additively growing with the model complexity. Ideally for 100 equidistant values and completely discovered goal function the accuracy is of the order of 1 %. Realistic expectations are far less favorable. Neither there are long enough data series that are without large chaotic behavior nor are any linearity in the goal data distribution. Widespread chaotic behavior lowers prediction accuracy while lowering model correlation. Ranking operation is insensitive to irregularities in data scales, but they differ significantly in value changes. Predicting from ranks is much more accurate for linear case. Still predictions of smaller data series (>25 data series) can be expected with about 10% accuracy which can be favorable for many practical applications on the web. There remains the task of more formal proof of theorem 1 and the elaboration of correlation in prediction error of single-variable systems. Errors in such systems are not statistically independent in respect to variable sampling and analytical operations.

REFERENCES

[1] Hamming, R. W., “Coding and Information Theory”, Prentice-Hall, SE, 1986

[2] Jović, F., “Qualitative Reasoning and a Circular Information Processing Algebra”, Informatica, 21, pp. 31-47, 1997

[3]Smith J.A. et al., “Method of airplane performance estimation and prediction”, US Patent 5,606,505, Feb 1997

[4] Wojsznis W.K., “Variable horizon predictor...”, US Patent 5,568,378, Oct 1996.

[5] Fox F.D. et al., “System and Method for the Advanced Prediction...”, US Patent 5,521,813 , May 1996
[6] Koza J.R., “Non-linear Genetic Algorithms for Solving Problems by Finding a Fit Composition of Functions”, PCT/US91/01970 , March 25, 1991
[7] Diambra L., Plastino A., “Modelling Time Series Using Information Theory”, Physical Letters A 216, pp. 278-282, 1996
[8] Perlovsky L.I., “Toward Physics of the Mind: Concepts, Emotions, Consciousness, and Symbols”, Physics of Life 3, 23-55, 2006
[9] Petz, B., “Osnovne statističke metode za nematematičare”, *Sveučilišna naklada Liber*, Zagreb, 1985
[10] Jović, F., “A Circular Qualitative Algebra”, Casys, 8:213-225, 2001
[11] 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
[12] Pelin, D., Fischer, D., Flegar, I., “Observing Chaos in a Ferroresonant Series Circuit”, International Conference on Electrical and Electronics Engineering, Turkey, December, 1999