Associative Model for the Forecasting of Time Series Based on the Gamma Classifier

Itzamá López-Yáñez^{1,2}, Leonid Sheremetov¹, and Cornelio Yáñez-Márquez³

¹ Mexican Petroleum Institute (IMP),
Av. Eje Central Lázaro Cárdenas Norte, 152, Mexico City, Mexico
{itzamal,sher}@imp.mx

² Instituto Politécnico Nacional,
Centro de Investigación y Desarrollo Tecnológico en Cómputo (CIDETEC - IPN),

Av. Juan de Dios Bátiz s/n Edificio CIDETEC, Mexico City, Mexico

³ Instituto Politécnico Nacional, Centro de Investigación en Computación
(CIC - IPN), Av. Juan de Dios Bátiz s/n Edificio CIC, Mexico City, Mexico
cyanez@cic.ipn.mx

Abstract. The paper describes a novel associative model for the forecasting of time series in petroleum engineering. The model is based on the Gamma classifier, which is inspired on the Alpha-Beta associative memories, taking the alpha and beta operators as basis for the gamma operator. The objective is to reproduce and predict future oil production in different scenarios in an adjustable time window. The distinctive features of the experimental data set are spikes, abrupt changes and frequent discontinuities, which considerably decrease the precision of traditional forecasting methods. As experimental results show, this classifier-based predictor exhibits competitive performance. The advantages and limitations of the model, as well as lines of improvement, are discussed.

Keywords: Time series forecasting, associative models, oil production time series, Gamma classifier.

1 Introduction

Time series (TS) analysis has become a relevant tool for understanding the behavior of different processes, both naturally occurring and human caused [1]. One of the latter kind of processes is the study of oil production through time, more specifically in fractured oil reservoirs, given their non-homogenous nature [2]. One of the tasks involved in such TS analysis is the prediction of future values (also known as TS forecasting), which is of particular interest in the context of industrial processes.

Computational Intelligence and Machine Learning have become standard tools for modeling and prediction of industrial processes in recent years, contributing models related mainly to Artificial Neural Networks (ANN) [3,4]. On the other hand, classical approaches, such as the Box-Jenkins Auto-Regressive Moving Average (ARMA) models, are still widely in use [1]. However, TS which exhibit non-linear, complex behavior tend to pose difficulties to such methods.

In the current paper, a novel associative model for the forecasting of TS is proposed. This method is based on the Gamma classifier (GC) [8], which is a supervised learning pattern classifier of recent emergence, belonging to the associative approach to Pattern Recognition. The GC has been previously applied to forecast atmospheric pollution TS [8], exhibiting a quite promising performance. However, that manner of forecasting had some limitations. In particular, it was able to predict only the following sample, given a known section of the TS.

This paper extends the previous work in order to predict a whole fragment of a TS, as well as to forecast samples located towards the past of a known fragment, allowing for a more complete reconstruction. The proposed method is applied to the prediction of several TS of oil production at a Mexican fractured oil reservoir, as well as to the synthetically generated Mackey-Glass TS, commonly used as a benchmark for TS forecasting.

The rest of the paper is organized as follows. Section 2 describes the TS used, while section 3 presents the GC, which is the basis for the proposal. The method presented here is further discussed in section 4, while section 5 introduces the experimental results and their discussion, and the conclusions and future work are included in the final section.

2 Oil Production Time Series

Since the beginning of petroleum engineering history, it has been a concern to forecast how long the production will last, sparking an ongoing interest for methods to predict production [4]. Yet, forecasting oil production is not an easy task. Production TS have several distinctive features regarding their predictability, which separate them from financial TS or physical processes TS, which are usually used for forecasting competitions [5].

First, monthly production TS are rather short: for a 30 years old brown oil-field, the longest series have about 300 data points; the rest of the series are shorter covering only several years or even months of production, so the data sets may only be a few dozens points long. Moreover, many TS are discontinuous: there are periods when for some reason the well is shut. Usually, a data set is predictable on the shortest time scales, but has global events that are harder to predict (spikes). These events can be caused by workovers, change of producing intervals and formations and so on. Some of these features (e.g. spikes, abrupt changes, discontinuities, and different trend segments) are evident in the plots of the TS used for experimentation, shown in figures 1 and 2. As can be seen, such TS lack typical patterns such as periodicity, seasonality, or cycles. Even though direct measurements are possible, the sampling process for these data is noisy.

Two data sets were used for the experiments: the first consisting of 6 randomly selected TS related to the monthly oil production and 6 wells in a Mexican oil-field, while the second is a synthetically generated TS coming from Mackey-Glass delay differential equations (MG equation) [6].

Table 1 summarizes statistical characteristics of the selected TS belonging to the first data set. TS forecasting is usually made on a one-year basis. Though the

Table 1. Statistical characteristics of the 6 selected TS of the first data set

Characteristic	TS1	TS2	TS3	TS4	TS5	TS6
Data points	154	296	126	131	139	140
Linear trend	$-264.52x$	N/A	$-84.437x$	$-0.0927x$	$-645.83x$	$-763.69x$
R^2	+79625		+89559	+9590.6	+96228	+111875
	0.27	N/A	0.02	0.00	0.67	0.43
Mean	59,366.91	46,083.38	84,440.67	6,692.23	58,381.15	55,221.67
Std. deviation	24,245.14	50,872.73	21,880.89	2,118.10	29,912.79	45,683.36
Mode	86,185.58	14,819.24	97,495.00	6,824.65	81,493.00	72,007.00
Skewness	-0.81	1.11	-0.58	0.57	-0.14	0.68
Kurtosis	-0.34	-0.31	2.64	1.67	-1.07	-0.10
Normality	0.89	0.87	0.92	0.92	0.95	0.90
Partial ACF	0.90	0.98	0.76	0.65	0.91	0.96

selected data set covered in all cases more than 10 years periods, discontinuities found almost in all TS substantially reduced the training basis.

To check for normality, the Shapiro-Wilk test was applied [7]. As can be seen from table 1, the normality test statistic (W) values for TS1 and TS2, though close but do not pass the test (W should be above 0.90 for normality). Many models used in TS analysis, including ARMA models, assume stationarity. To check for stationarity both parametric (partial autocorrelation function, ACF) and nonparametric (runs) tests were applied. Though nonparametric tests for stationarity are more applicable for our dataset since TS1 and TS2 do not meet the normality assumption, the auto-correlation does seem to agree with results of the nonparametric runs test. Partial ACF plots (calculated for 18 lags) show similar results for most of the wells. Most of the autocorrelation is actually just a lag 1 (1 month) effect (shown in table 1), with minor effects at other lags (maximum of 0.32 for lag of 6 months for TS6). The results indicate that there is actually little pattern in the data —though having a distinct decline trend, they have very little periodicity, and only very short term, month-to-month correlation between observations.

3 Gamma Classifier

This supervised learning associative model was originally designed for the task of pattern classification, and borrows its name from the operation at its heart: the generalized Gamma operator [8,9]. This operation takes as input two binary patterns — $\mathbf{x}$ and $\mathbf{y}$ — and a non-negative integer θ and gives a 1 if both vectors are similar (at most as different as indicated by θ) or 0 otherwise. Given that the Gamma operator uses some other operators (namely α , β , and u_β), they will be presented before. The rest of this section is strongly based on [8,9].

Definition 1 (Alpha and Beta operators). *Given the sets $A = \{0,1\}$ and $B = \{0,1,2\}$, the alpha (α) and beta (β) operators are defined in a tabular*

form as shown in table 2. The corresponding vector versions of both operators for inputs $\mathbf{x} \in A^n$, $\mathbf{y} \in A^n$, and $\mathbf{z} \in B^n$ give an n -dimensional vector as output, whose i -th component is computed as follows.

$$\alpha(\mathbf{x}, \mathbf{y})_i = \alpha(x_i, y_i) \quad \text{and} \quad \beta(\mathbf{z}, \mathbf{y})_i = \beta(z_i, y_i) \quad (1)$$

Definition 2 (u_β operator). Considering the binary pattern $\mathbf{x} \in A^n$ as input, this unary operator gives the following integer as output.

$$u_\beta(\mathbf{x}) = \sum_{i=1}^n \beta(x_i, x_i) \quad (2)$$

Definition 3 (Gamma operator). The similarity Gamma operator takes two binary patterns — $\mathbf{x} \in A^n$ and $\mathbf{y} \in A^m$; $n, m \in \mathbb{Z}^+$ $n \leq m$ — and a non-negative integer θ as input, and outputs a binary number, according to the following rule.

$$\gamma_g(\mathbf{x}, \mathbf{y}, \theta) = \begin{cases} 1 & \text{if } m - u_\beta[\alpha(\mathbf{x}, \mathbf{y}) \bmod 2] \leq \theta \\ 0 & \text{otherwise} \end{cases} \quad (3)$$

where $\bmod$ denotes the usual modulo operation.

The GC makes use of the previous definitions, as well as that of the Modified Johnson-Möbius code [10] in order to classify a (potentially unknown) test pattern $\tilde{\mathbf{x}}$, given a fundamental set of learning or training patterns $\{(\mathbf{x}^\mu, \mathbf{y}^\mu)\}$. To do this, it follows the algorithm described below.

1. Convert the patterns in the fundamental set into binary vectors using the Modified Johnson-Möbius code.
2. Code the test pattern with the Modified Johnson-Möbius code, using the same parameters used for the fundamental set.
3. Compute the stop parameter $\rho = \bigwedge_{j=1}^n \bigvee_{i=1}^p x_j^i$.
4. Transform the index of all fundamental patterns into two indices, one for their class and another for their position in the class (e.g. $\mathbf{x}^\mu$ in class i becomes $\mathbf{x}^{i\omega}$).

Table 2. The alpha (α) and beta (β) operators

$\alpha : A \times A \rightarrow B$			$\beta : B \times A \rightarrow A$		
x	y	$\alpha(x, y)$	x	y	$\beta(x, y)$
0	0	1	0	0	0
0	1	0	0	1	0
1	0	2	1	0	0
1	1	1	1	1	1
			2	0	1
			2	1	1

5. Initialize $\theta = 0$.
6. Do $\gamma_g(x_j^{i\omega}, \tilde{x}_j, \theta)$ for each component of the fundamental patterns.
7. Compute a weighted sum c_i for each class, according to this equation:

$$c_i = \frac{\sum_{\omega=1}^{k_i} \sum_{j=1}^n \gamma_g(x_j^{i\omega}, \tilde{x}_j, \theta)}{k_i} \quad (4)$$

where k_i is the cardinality in the fundamental set of class i .

8. If there is more than one maximum among the different c_i , increment θ by 1 and repeat steps 6 and 7 until there is a unique maximum, or the stop condition $\theta \geq \rho$ is fulfilled.
9. If there is a unique maximum among the c_i , assign $\tilde{\mathbf{y}}$ to the class corresponding to such maximum.
10. Otherwise, assign $\tilde{\mathbf{y}}$ to the class of the first maxima.

The main characteristic of the GC, which sets it apart from other classifiers, is that the measure of similarity between patterns (on which the classification decision is based) is computed independently on each feature, and later integrated for the whole pattern. This fundamental difference allows this model to offer better performance than more conventional models on those problems for which computing the similarity using all features give rise to ambiguity. This property is useful to address such issues as spikes and abrupt changes: if the majority of features are similar enough among two patterns, the differences between two particular features have little bearing on the final outcome. Another defining and desirable characteristic of the GC is its low computational complexity, which is $O(pn)$ for a fundamental set of p n -dimensional learning patterns [8,9].

4 Proposed Model

As mentioned before, the GC has been already applied to forecast TS, but predicting only the following sample. The current proposal takes the previous work and extends it in order to predict not only the first unknown sample in a fragment of a TS, but the whole fragment (of arbitrary length). Also, it is now possible to forecast samples located towards the past of a known fragment (i.e. towards the left of the TS, or previous to the known segment), which allows for a more complete reconstruction. In order to achieve the former objectives, the coding and pattern building method introduced in [8] is generalized in order to consider negative *separations* between input and output patterns, as well as separations greater than one sample away.

Definition 4 (Separation). *Given a TS D with samples $d_1 d_2 d_3 \dots$, the separation s between a segment $d_i d_{i+1} \dots d_{n-1}$ (of length n) and sample d_j is given by the distance between the closest extreme of the segment and the sample.*

Example 1. Let D be the TS with the following sample values: $D = 10, 9, 8, 7, 6, 5, 4, 3, 2, 1$. Considering the segment D_1 of size 3 as $D_1 = d_3 d_4 d_5 = 8, 7, 6$ then sample $d_6 = 5$ is at a separation $s = 1$, sample $d_7 = 4$ is at $s = 2$, sample $d_{10} = 1$ is at $s = 5$, and sample $d_1 = 10$ is at $s = -2$.

Based on this definition and the GC, the proposed TS forecasting method follows the algorithm presented below, considering a TS D of length l with a prediction (test) segment of length t , and a length for the patterns of size n .

1. Starting from the TS D , calculate the differences between successive samples in order to work with values relative to that of the previous sample. The new time series D' has length $l - 1$.

$$D_{l \times 1} \longrightarrow D'_{(l-1) \times 1} \quad (5)$$

2. Build a set of associations between TS difference segments of length n and its corresponding difference with separation s , for both positive and negative values of $s = 1, 2, \dots, t - 1, t$.

Thus, there will be $2t$ sets of associations of the form $\{\mathbf{a}^\mu, \mathbf{b}^\mu\}$ where $\mathbf{a} \in \mathbb{R}^n$ and $\mathbf{b} \in \mathbb{R}$, and the i -th association of the s -th set is made up by $\mathbf{a}^\mu = d_i d_{i+1} \dots d_{i+n-1}$ and:

$$\mathbf{b}^\mu = \begin{cases} d_{i+n+s-1} & \text{if } s > 1 \\ d_{i-s} & \text{if } s < 1 \end{cases} \quad (6)$$

3. Train a different GC from each association set; there will be $2t$ different classifiers, each with a distinct fundamental set $\{\mathbf{x}^\mu, \mathbf{y}^\mu\}$ for its corresponding value of s .
4. Operate each GC with all the input segments $\mathbf{a}^\mu$.
5. When multiple classifiers give different output values $\tilde{\mathbf{y}}$ for the same data point in the differences TS D' , there are two prominent alternatives to integrate them into one value $\tilde{\mathbf{y}}'$.
 - (a) Average the values given by the two classifiers with the same absolute separation $|s| = \{-s, s\}$; this is denoted as the *combined* method.
 - (b) Average the values given by all available classifiers; this is known as the *combined average* method.
6. Convert back to absolute values by adding the forecast relative value ($\tilde{\mathbf{y}}'$) to the original value of the previous sample, taken from D .

$$\tilde{\mathbf{y}}_i = D_{i-1} + \tilde{\mathbf{y}}'_i \quad (7)$$

Thus, the proposed method uses a classifier to obtain a particular value for each fundamental set (associated to a different separation), in order to later integrate these output values into a single value, which may or may not be known. With this method, a TS with no discontinuities will be reconstructed and predicted in its entirety, except for the first sample. This is due to the differences taken as the first step, which decreases the length of TS D' in one with respect to the original series D .

Also, given the guaranteed correct recall of known patterns exhibited by the GC [8], it is most likely that the known segments of the TS will be exactly reconstructed (i.e. error of 0). There will be a difference between the original and forecast values only if the same input pattern ($\mathbf{a}^i = \mathbf{a}^j$) appears more than once, associated to different output values ($\mathbf{b}^i \neq \mathbf{b}^j$).

$$\exists (\mathbf{a}^i, \mathbf{b}^i) \wedge (\mathbf{a}^j, \mathbf{b}^j), i \neq j \text{ such that } \mathbf{a}^i = \mathbf{a}^j \wedge \mathbf{b}^i \neq \mathbf{b}^j \quad (8)$$

5 Experimental Results

In order to test the proposed prediction model, it was applied to two data sets: one made up by several TS related to oil production, which were previously described, and the other a synthetically generated TS coming from the MG equation [6].

For comparison purposes, two optimality measures were used: the Mean Square Error (MSE) and the adjusted Mean Absolute Percent Error introduced by Armstrong, also known as symmetric MAPE (SMAPE, though it is not symmetric since over- and under-forecasts are biased). These error metrics are computed as shown in equations 9 and 10, where y_j is the actual and $\hat{y}_j$ is the predicted value, with N being the amount of samples considered.

$$E_{MSE} = \frac{1}{N} \sum_{i=1}^N (y_i - \hat{y}_i)^2 \quad (9)$$

$$E_{SMAPE} = \frac{1}{N} \sum_{i=1}^N \left| \frac{y_i - \hat{y}_i}{y_i + \hat{y}_i} \right| \quad (10)$$

5.1 Oil Production TS Experiment

For the first data set, three scenarios were considered, with test segments of 12, 18, and 24 samples of length, in order to find out how the length of the test segment affects prediction. Meanwhile, the experimental results are summarized in table 3, and for illustrating purposes, the resulting monthly oil production predictions for two TS are shown in figures 1 and 2.

Fig. 1. Prediction of TS 1

Fig. 2. Prediction of TS 2

As can be seen, the performance of the proposed method varies throughout the different experiments, as was expected given the differences presented by the TS used. TS 1 and 3 show very good results on both metrics, while TS 4 has very good MSE and reasonably good SMAPE. On the other hand, the metrics for TS 5 are large for both metrics. When looking at the plots, the reasons for such behaviors start to emerge. TS 1 has no discontinuities, while out of the 296 samples of TS 2, only 146 are valid data points grouped in 5 segments. Under

Table 3. Experimental results of oil production data for Mean Square Error in millions ($\text{MSE} \times 1.0\text{E}+06$) and Symmetric Mean Absolute Percentage Error (SMAPE)

Mean Squared Error						
Time Series	Combined			Combined Avg.		
	$l=12$	$l=18$	$l=24$	$l=12$	$l=18$	$l=24$
TS 1	24.01	27.50	28.82	20.87	20.96	36.08
TS 2	33.76	38.11	105.28	48.39	64.32	60.49
TS 3	8.71	20.30	35.59	8.71	28.78	36.55
TS 4	0.99	3.43	4.06	1.19	2.60	5.62
TS 5	297.13	195.10	206.62	297.13	291.22	202.31
TS 6	28.05	64.07	4.12	29.48	82.83	15.08
Symmetric Mean Absolute Percentage Error						
Time Series	Combined			Combined Avg.		
	$l=12$	$l=18$	$l=24$	$l=12$	$l=18$	$l=24$
TS 1	0.0304	0.0349	0.0323	0.0304	0.0272	0.0366
TS 2	0.2252	0.3172	0.2908	0.3162	0.3172	0.2908
TS 3	0.0182	0.0248	0.0284	0.0182	0.0264	0.0328
TS 4	0.0508	0.1376	0.1046	0.0488	0.1140	0.1609
TS 5	0.1858	0.1321	0.1334	0.1858	0.1699	0.1215
TS 6	0.4329	0.3893	0.0711	0.3616	0.3574	0.1516

such conditions, it can only be expected that any predictor will have a hard time obtaining good results.

It is noteworthy that for all six TS, on both variants (*Combined* and *Combined average*) and both performance metrics, the best results were obtained with test segments of length $l = 12$. It is well-known in TS forecasting that the farther away from the learning samples, the greater the uncertainty in the prediction [1]. However, given the characteristics and properties of the GC (i.e. a segment and a sample with arbitrary separation could be associated and learned) the effect of l was previously unknown.

Another interesting finding is that some TS yield quite good results under some performance metrics, while the results for the other metric are surprisingly poor. Such is the case of TS 6, whose MSE ($1.51\text{E}+07$ best, $3.73\text{E}+07$ average) is below the mean of the six TS ($6.61\text{E}+07$), while its SMAPE (0.2940 average, 0.4329 worst) is considerably above the mean (0.1502 , respectively). Yet, each metric emphasizes different aspects of the prediction error, thus giving rise to such apparent discrepancies. What remains to be found is why such large inconsistencies in performance arise.

5.2 Mackey-Glass TS Experiment

On the other hand, the TS for the second data set is generated by the MG equation, which has been extensively used by different authors to perform comparisons between different techniques for forecasting and regression models. The MG equation is explained by the time delay differential equation defined as:

Table 4. Comparison of performance on the Mackey-Glass benchmark [6,11]

Model	RMSE
Gamma Classifier 0.001502	
MLMVN	0.0063
CNNE	0.009
SuPFuNIS	0.014
GEFREX	0.0061
EPNet	0.02
GFPE	0.026
Classical BP NN	0.02
ANFIS	0.0074

$$y(t) = \frac{\alpha y(t - \tau)}{1 + y^c(t - \tau)} - \beta y(t) \quad (11)$$

where α , β , and c are parameters and τ is the delay time. According as τ increases, the solution turns from periodic to chaotic. To make the comparisons with earlier work, one thousand data points are generated with an initial condition $x(0) = 1.2$ and $\tau = 17$ based on the fourth-order Runge-Kutta method with time step 0.1. The following parameters (taken from consulted literature) have been used as a benchmark: $\alpha = 0.2$, $\beta = 0.1$, $c = 10$ [6]. From these 1000 samples, the first 500 were used for training and the latter 500 points were used for testing. Rooted MSE (RMSE, the square root of the MSE) was used given that it is the performance measure of choice for this TS through the consulted literature, obtaining a RMSE of 0.001502. Table 4 includes a comparison against other models [6,11].

As can be seen, the proposed model based on the GC clearly outperforms the other techniques, giving an RMSE which is approximately four times smaller than that of GEFREX (0.001502 against 0.0061), its closest competition.

6 Conclusions and Future Work

In this paper the forecasting capabilities of an associative model are tested. The main result is that a pattern classifier is successfully applied to the task of TS prediction, though it is not the kind of tasks envisioned while designing and developing the GC. Moreover, the proposed method uses only one variable when there are more available, and the phenomenon under study is considered to be a rather complex one. Experiments with the GC-based predictor under multivariate settings for oil-production forecasting are on-going. One of such scenarios is the development of a hierarchichal model, which begins by detecting and classifying specific events, in order to use a specific predictor.

On a different matter, the dependencies of forecasting uncertainty on separation between predicted data point and known data have been confirmed, despite

the independency offered by the proposed model. This indicates a likely dependency in the actual data sampled from the natural phenomenon, which remains to be tested.

Acknowledgements. This research was partially supported by the CONACYT-SENER (project 146515), as well as the Instituto Politécnico Nacional (Secretaría Académica, COFAA, SIP, CIDETEC and CIC), CONACyT, and SNI.

References

1. Schelter, B., Winterhalder, M., Timmer, J. (eds.): Handbook of Time Series Analysis. Wiley, Weinheim (2006)
2. van Golf-Racht, T.D.: Fundamentals of Fractured Reservoir Engineering, Developments in Petroleum Science, vol. 12. Elsevier, Amsterdam (1982)
3. Palit, A.K., Popovic, D.: Computational Intelligence in Time Series Forecasting. Springer, London (2005)
4. Sheremetov, L., Alvarado, M., Bañares-Alcántara, R., Anminzadeh, F.: Intelligent Computing in Petroleum Engineering (Editorial). J. of Petroleum Science and Engineering 47(1-2), 1–3 (2005)
5. He, Z., Yang, L., Yen, J., Wu, C.: Neural-Network Approach to Predict Well Performance Using Available Field Data. In: SPE Western Regional Meeting, Bakersfield, California, March 26-30, SPE 68801 (2001)
6. Kim, D., Kim, C.: Forecasting Time Series with Genetic Fuzzy Predictor Ensemble. IEEE Tr. on Fuzzy Systems 5(4), 523–535 (1997)
7. Johnson, R.A., Wichern, D.W.: Applied Multivariate Statistical Analysis (Prentice Hall series in statistics), 5th edn. Prentice Hall (2001)
8. López-Yáñez, I., Argüelles-Cruz, A.J., Camacho-Nieto, O., Yáñez-Márquez, C.: Pollutants Time-Series Prediction Using the Gamma classifier. Int. J. of Computational Intelligence Systems 4(4), 680–711 (2011)
9. Acevedo-Mosqueda, M.E., Yáñez-Márquez, C., López-Yáñez, I.: Alpha-Beta Bidirectional Associative Memories: Theory and Applications. Neural Processing Letters 26(1), 1–40 (2007)
10. Yáñez, C., Felipe-Riveron, E., López-Yáñez, I., Flores-Carapia, R.: A Novel Approach to Automatic Color Matching. In: Martínez-Trinidad, J.F., Carrasco Ochoa, J.A., Kittler, J. (eds.) CIARP 2006. LNCS, vol. 4225, pp. 529–538. Springer, Heidelberg (2006)
11. Aizenberg, I., Moraga, C.: Multilayer feedforward neural network based on multi-valued neurons (MLMVN) and a backpropagation learning algorithm. Soft Computing 11, 169–183 (2007)