Comparative Analysis of Real and Binary Coded Genetic Algorithm for Fuzzy Time Series Prediction

Shilpa Jain* Dinesh Bisht† Prakash C. Mathpal‡

Received: 04 September 2013 | Revised: 08 October 2013 | Accepted: 28 October 2013 | Available online: 01 December 2013

Keywords
- Genetic Algorithms; computational intelligence
- Micro GA, Messy GA, Sawtooth GA, Differential Evolution GA
- Real coded GA, binary coded GA, forecasting of enrollments of University of Alabama

Abstract
Genetic Algorithm (GA) form a subset of evolutionary computing, rapidly growing area of AI (more particularly computational intelligence). Some variants of GA are binary coded GA, real coded GA, micro GA, messy GA, sawtooth GA, differential evolution GA. Here, in this research article we present a comparative analysis of real coded and binary coded GA for forecasting of enrollments of University of Alabama. Results clearly state that real coded GA are faster & more accurate. Results are stated after testing on 100 generations.

Introduction
This Genetic Algorithm (GA):- GA models genetic evolution, invented by John Holland [1] in 1975. GA is population based, probabilistic search and optimization technique that works on mechanism of natural genetics and Darwin's principle of natural selection (i.e. survival of the fittest).

GA maintains the population of individuals that evolves according to rules of selection and manipulates by genetic operators - crossover and mutation. GA can search irregular space and hence are applied to a variety of function optimization, parameter estimation and machine learning applications.

Working principle[2] can be explained briefly using the flowchart.
The primary difference between binary and real coded GA is that in binary coded GA variables are represented by bits of zeros and ones while GAs based on real number representation are called real coded GAs (RCGAs).

GA originated with a binary representation of the variables. Until 1991 no specific theoretical studies were done for RCGAs so its use was controversial. RCGAs are mainly used for numerical optimization on continuous domains (Wright[4]; Davis[5]; Mchalewicz[6]; Michlenbein et al.[7]; Herrera et al.[8]). Goldberg[9], Deb[10-12] are also eminent researchers who contributed in development of crossover and mutation operators of RCGAs.

In this research article we compare real coded GA with binary coded GA implemented on data set of University of Alabama which is a fuzzy time series. Time series is a sequence of data points recorded regularly. Fuzzy Time Series is used which differs from traditional time series as values of Fuzzy Time Series are represented by Fuzzy sets rather than real values. Song and Chissom proposed definitions of Fuzzy Time Series [13-15]. Fuzzy time series has been developing for three decades to improve forecasting accuracy of enrollments, economy, stock market, weather, population growth etc.

**Methodology**

Proposed method for Real and binary coded GA fuzzy time series forecasting:

1. Define the universe of discourse.
2. Partition into intervals using G.A (For binary coded GA variables are population strings of zeros and ones and for real coded GA population strings are real number representations).
3. Construct Fuzzy sets.
4. Fuzzify the data.
5. Established Fuzzy rules.
6. Forecast.
7. Forecasting accuracy is measured using MSE. The lower the MSE, the better is forecasting method. MSE is defined by the expression

\[
\text{MSE} = \frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2
\]
\[ \text{MSE} = \frac{\sum_{i=1}^{n} (\text{actual}_i - \text{forecasted}_i)^2}{n} \]

Implementation of Proposed Method to Enrollments forecast

Step 1:- Define the Universe of discourse, U based on range of available historical time series data, by rule U=[Dmin - D1, Dmax + D2] Where, D1 and D2 are two proper positive numbers. Table 1 shows historical data of enrollments for University of Alabama. From table 1, we can see Dmin = 13055 and Dmax = 19337. Let D1 = 55 and D2 = 663, therefore the universe of discourse U = [13000, 20000]

Step 2:- Divide universe of discourse U into 21 intervals using X1 to X19.

Table 1: Data of Enrollments from University of Alabama.

<table>
<thead>
<tr>
<th>Year</th>
<th>Actual Enrollments</th>
</tr>
</thead>
<tbody>
<tr>
<td>1971</td>
<td>13055</td>
</tr>
<tr>
<td>1972</td>
<td>13563</td>
</tr>
<tr>
<td>1973</td>
<td>13867</td>
</tr>
<tr>
<td>1974</td>
<td>14696</td>
</tr>
<tr>
<td>1975</td>
<td>15460</td>
</tr>
<tr>
<td>1976</td>
<td>15311</td>
</tr>
<tr>
<td>1977</td>
<td>15603</td>
</tr>
<tr>
<td>1978</td>
<td>15861</td>
</tr>
<tr>
<td>1979</td>
<td>16807</td>
</tr>
<tr>
<td>1980</td>
<td>16919</td>
</tr>
<tr>
<td>1981</td>
<td>16388</td>
</tr>
<tr>
<td>1982</td>
<td>15433</td>
</tr>
<tr>
<td>1983</td>
<td>15497</td>
</tr>
<tr>
<td>1984</td>
<td>15145</td>
</tr>
<tr>
<td>1985</td>
<td>15163</td>
</tr>
<tr>
<td>1986</td>
<td>15984</td>
</tr>
<tr>
<td>1987</td>
<td>16859</td>
</tr>
<tr>
<td>1988</td>
<td>18150</td>
</tr>
<tr>
<td>1989</td>
<td>18970</td>
</tr>
<tr>
<td>1990</td>
<td>19328</td>
</tr>
<tr>
<td>1991</td>
<td>19337</td>
</tr>
<tr>
<td>1992</td>
<td>18876</td>
</tr>
</tbody>
</table>

Step 3:- Define each chromosome consisting of 19 genes as shown below:-

Table 2: Chromosome consisting of 19 genes

| X1 | X2 | X3 | .... | .... | .... | .... | X19 |

Where X1, X2, X3,... are integer variables. These values of X1, X2, X3,... X19 are generated randomly by GA.
GA generates 30 chromosomes as initial population in form of binary strings for binary coded GA and in form of real number representation for RCGAs. Fuzzify the historical enrolments shown in Table 1 with the chromosomes in the population. Let the Fuzzy sets constructed can be realized as

\[ A_1 = [13000, X_1, X_2] \]
\[ A_2 = [X_3, X_4, X_5] \]
\[ A_3 = [X_6, X_7, X_8] \]
\[ A_4 = [X_9, X_{10}, X_{11}] \]
\[ A_5 = [X_{12}, X_{13}, X_{14}] \]
\[ A_6 = [X_{15}, X_{16}, X_{17}] \]
\[ A_7 = [X_{18}, X_{19}, 20000] \]

Let the Fuzzy set \( A_1, A_2, \ldots, A_7 \) be linguistic values of the linguistic variable enrollments shown as below:

\( A_1 = \) EL (Extremely Low) \( A_2 = \) VL (Very Low) \( A_3 = \) L (Low) \( A_4 = \) M (Medium) \\
\( A_5 = \) H (High) \( A_6 = \) VH (Very High) \( A_7 = \) EH (Extremely High)

**Step 4:** Rules for forecasting

\[ A_1 \rightarrow A_2 \]
\[ A_2 \rightarrow A_2, A_2 \rightarrow A_3 \]
\[ A_3 \rightarrow A_3, A_3 \rightarrow A_4 \]
\[ A_4 \rightarrow A_4, A_4 \rightarrow A_6 \]
\[ A_5 \rightarrow A_3 \]
\[ A_6 \rightarrow A_5, A_6 \rightarrow A_6, A_6 \rightarrow A_7 \]
\[ A_7 \rightarrow A_7 \]

**Fig. 3 Rule Base**

The rules stated in Figure 3 are the standard rules given by Chen [16].

**Step 5:** In this research article we have used Matlab 7.6.0 to code Fuzzy and GA techniques of soft computing. Code is generated for Fuzzy using FIS and command that is used for the same is EVALFIS(). \( Y = \text{EVALFIS}(U, \text{FIS}) \) simulates the Fuzzy Inference System FIS for the input data \( U \) and returns the output data \( Y \). For a system with \( N \) input variables and \( L \) output variables,

* \( U \) is a \( M \)-by-\( N \) matrix, each row being a particular input vector
* Y is M-by-L matrix, each row being a particular output vector.

Step 6:- MSE for each model of each generation is calculated which becomes the objective function for GA.

Step 7:- Selection rate is 50 percentage. So that means rest 50 percentage of the chromosomes are discarded and new chromosomes after performing cross over and mutation operations are added to the population. This continues for hundred generations.

**Results & Conclusion**

In this research article we have proposed a comparative analysis of binary and real coded GA. The method was implemented on the historical time series data of student enrollment at the University of Alabama to provide comparative study. When the mutation rate is 0.05, the crossover is 1, the number of intervals in the universe of discourse is 19, the population size=30, and the number of generations evolving is 100 for both binary and real coded GA. GA originated with a binary representation of the variables. When variables are quantized naturally, the binary GA can be applied. However, when the variables are continuous, it's more logical to represent them by floating point numbers. In addition, since the binary GA has its precision limited to binary representation of variables using floating point numbers that allows representation to machine precision. Thus continuous GA or real coded GA also has advantage of requiring less storage than binary GA because a single floating point number represents the variable instead of Nbits integers.

Moreover the continuous GA is inherently faster than the binary GA because the chromosome don’t have to be decoded prior to evaluation of objective function.

**Table 3: shows a comparison of the MSE of the forecasting enrollments of the binary and real coded GA.**

<table>
<thead>
<tr>
<th>Model</th>
<th>Real coded GA</th>
<th>Binary Coded GA</th>
</tr>
</thead>
<tbody>
<tr>
<td>MSE</td>
<td>395420</td>
<td>275900</td>
</tr>
</tbody>
</table>

With the above specifications it was observed that Real coded GA program ran much faster in comparison to Binary coded GAs although accuracy in terms of MSE is better for binary coded GA although as suggested by Herrera, Lozano & Verdegay [17] results can be further improved by using non-uniform mutation and BLX-α, logical FCB and linear crossover operators.

**References**