International Journal of Soft Computing 2 (5): 612-616, 2007
ISSN: 1816-9503

Modeling Monthly Mean Maximum Temperature Using Genetic Programming

[1]S. Shahid, [1]M. Hasan and [2]R.U. Mondal
[1]Department of Applied Physics and Electronics,
[2]Department of Information and Communication Technology,
Rajshahi University, Rajshahi-6205, Bangladesh

Abstract: Climate is a continuous, data-intensive, multidimensional, dynamic and chaotic process. Conventional classical methods, generally used for prediction from historical time series data, often fail to predict climate reliably. Recently, various soft computing techniques are being used for their prediction. Genetic programming has been used in this study for the modeling of monthly mean maximum temperature. The result is compared to that obtained by using neural network. The study shows that the model produced by genetic programming can be used for the reliable prediction of monthly mean maximum temperature of the area. Though the result obtained by genetic programming is more erroneous compared to neural network, it provides an equation that could be used for reasonable prediction of monthly mean maximum temperature manually.

Key words: Climate prediction, monthly mean maximum temperature, time series modeling, genetic programming, artificial neural network

INTRODUCTION

Climate is a continuous, data-intensive, multidimensional, dynamic and chaotic process. The parameters that are required to predict the climatic elements are enormously complex and subtle so that uncertainty in a prediction using all these parameters is large. Furthermore, it often happens that there is not even a reasonable qualitative understanding of the phenomenon, or the phenomenon is the result of complex interactions involving many independent and irreducible degrees of freedom. In those cases all the conventional methods fail to predict weather with reasonable accuracy. Statistical methods are being used for a long time in such cases.

In statistical forecasting systems, predictable climatic element is represented as a sequence of vector depending on time. The forecasting is viewed as a problem of function approximation, where a method is used to approximate the continuous valued function as closely as possible. A number of statistical methods are available that can be applied for this purpose. Although these statistical methods are very useful and have been utilized for many years for predictions, they are somewhat limited in their ability to forecast in certain situations. For example, they often found to fail in dealing with ill-defined and uncertain systems. Recently, Artificial Intelligence (AI) based techniques are being used to overcome these situations. It has been found that due to the adaptability, intrinsic parallelism and efficiency in solving complex problems, AI techniques are much more efficient in predicting weather time series compared to conventional statistical methods (Philip and Joseph, 2001).

Among the AI techniques, so far mostly Artificial Neural Networks (ANNs) have been used for the prediction of weather time series data (Maqsood *et al*., 2000a, 2000b; Mondal and Shahid, 2003). Recently, Genetic Programming (GP) is also being used for the modeling of chaotic time series (Zhang *et al*., 2004). Though a few researches have been carried out on the application of genetic programming in weather forecasting, it has been found very much promising. The advantage of GP is that unlike other AI based forecasting models, it does not use a pre-set forecasting model. GP has the potential to search for an appropriate model with optimal parameter values automatically. Kishtawal *et al*. (2003) studied the feasibility of a nonlinear technique based on genetic programming for the prediction of summer rainfall over India and successfully identified the equations that best describe the temporal variations of the seasonal rainfall over India. Coulibaly (2004) successfully used GP to simulate local scale daily extreme (maximum and minimum) temperatures based on large-scale atmospheric variables. He also compared the performance of the GP based models to a commonly used statistical model and showed that the models evolved by GP are simpler and more efficient than the common statistical methods. In this study, genetic programming is proposed

Corresponding Author: S. Shahid, Department of Applied Physics and Electronics, Rajshahi University, Rajshahi-6205, Bangladesh

for the simulation of monthly mean maximum temperature of Rajshahi city and its surrounding areas. The climate of the area is characterized by high temperature, heavy rainfall, often-excessive humidity and fairly marked seasonal variations (Rashid, 1979). Temperature is one of the dominant elements of the climate of the area.

GENETIC PROGRAMMING AND ITS APPLICATION IN TIME SERIES FORECASTING

Genetic Programming, a subset of genetic algorithms, is an automated methodology inspired by biological evolution to find computer programs that best perform a user-defined task. It is therefore, a particular machine learning technique that uses an evolutionary algorithm to optimize a population of computer programs according to a fitness landscape determined by a program's ability to perform a given computational task. The steps followed by genetic programming are given:

1. Generate an initial population of candidate solutions or programs. A candidate program is produced by randomly constructing a tree made up of operators and inputs.
2. Execute each program in the population over a set of training data and ranked based on its prediction error.
3. Perform mutation, crossover and other genetic operators on the selected individuals and form the new population using the result.
4. The best computer program that appeared in any generation is the best-so-far solution. If the solution is sufficient, end the process and present the best program in the population as the result. Otherwise go to step 2.

Genetic programming works best in finding solution of problem that have no ideal solution. One of the most interesting applications of genetic programming is the analysis of time series data, which is a result of complex interactions of many independent variables. Generally, there is no ideal equation that represents a time series. GP can be used to model the time series and find the best equation or program that represent the time series.

The first study of GP based time series forecasting was done by Koza (1992). His target was to predict only the next value of a time series produced by a logistic equation. A more difficult task was attempted by Oakley (1994). He studied the time series produced by Mackey-Glass equation for a non-standard long-term prediction. Iba (1995) introduced a new GP based approach for the solution of system identification and applied it for time series prediction problem. Like Oakley (1994), he also used the time series generated with the Mackey-Glass equation and solve the problem more efficiently. Besides these, a number of recent works have been carried out on short-term chaotic time series modeling and forecasting using GP (Zhang *et al*., 2004; Angline, 1998; Cortez *et al*., 2001).

THEORY OF TIME SERIES PREDICTION USING GP

The task of time series prediction is to find a model or function, f, which maps the future value from the past values. The input parameters of the model are current time series index, x_n and L past values of the time series, i.e., i.e., x_n, x_{n-1}, x_{n-2} ,...., x_{n-L}. The prediction model maps the future value, $\hat{x}_{n+1}$, from the past values,

$$\hat{x}_{n+1} := f(x_n, x_{n-1}, x_{n-2}, ..., x_{n-L}) \quad (1)$$

The model building consists of determining the structure and parameters of this mapping. GP solve this problem by randomly generating a set of models or individuals, each represents a tree. The tree representations consist of nodes and are of variable length. The nodes can either be non-terminal nodes consist of functions that perform some action on one or more signals within the structure to produce an output signal, or terminal nodes that represent an input variable or a constant. Fitness of each individual is computed by measuring its prediction ability. Prediction ability is measured by computing the Root Mean Square Error (RMSE) in predicting the future value from the past values of the given time series.

$$RMSE = \frac{1}{n}\sum_{i=1}^{n}(x_i - \hat{x}_i)^2 \quad (2)$$

After evaluating the fitness, better individuals are selected for the next generation. Genetic operators, viz. mutation and crossover are used to generate offspring form the existing individuals. Crossover between two trees is carried out by randomly choosing a branch in each tree and replacing them with each other. Mutation is performed by choosing a node and changing its value or meaning i.e., the function symbol could become another function symbol or be deleted and the terminal value could be modified. The process is repeated until the termination conditions are met. The fittest individual in the final generation is the prediction model of the given time series.

PREDICTION OF MONTHLY MEAN MAXIMUM TEMPERATURE USING GP

The goal of a GP monthly mean maximum temperature modeling process is to produce an algebraic expression or model that best describes the maximum temperature time series. For this purpose, the GP algorithm works with

solution candidates that are tree structure representations of symbolic expressions. The function set used for non-terminal nodes are, +, -, *, /, % and log. Here, / represents division which returns the result of division by 0.01 if the denominator is 0, otherwise, returns the actual result of division. As the goal is to find the future values from past values, the terminal nodes must consist of previous values of the time series as well as constants. However, choosing which past values to employ is difficult. As the number of the past values increases, the size of the search space increases and the search task become difficult. Therefore, the judicious choice of the maximum number of past values is important. In the present case, eight previous temperature values are chosen from the knowledge of other climatic models. Among the eight temperature values, four values are selected from present year and four are from previous year. The input set for terminal nodes are,

$$\left\{\begin{matrix} x_n, x_{n-1}, x_{n-2}, x_{n-3}, \\ x_{n-12}, x_{n-13}, x_{n-14}, x_{n-15}, \Re \end{matrix}\right\} \quad (3)$$

Where, $\Re$ are the random constant ranges from 0-10.0.

The performance of genetic programming depends on the parameter values used. Trial and error method is used to choose the best configuration of GP for the prediction of monthly mean maximum temperature, which is given:

Parameters	Values
Limit of maximum tree size	7
Crossover probability	0.8
Mutation probability	0.01
Population size	100
Maximum number of generation	250

RESULTS AND DISCUSSION

For the prediction of monthly mean maximum temperature, data for the months of January 1964 to February 1994 are used as the learning set of GP and the data for the months of March 1994 to June 2004 are used for validation. GP takes 123 epochs (165 sec) to give the lowest training error (RMSE = 1.4838). The convergence of the algorithm during training is shown in Fig. 1(a). The equation tree that produced the best result is shown in Fig. 1(b). The simplified equation that best predict the monthly mean maximum temperature data of the study area is,

$$\hat{x}_{n+1} = x_{n-12} - \frac{x_{n-13} - x_n}{3.7062 \dfrac{x_{n-14} - 5.4081}{3.6897 - x_{n-1}}} \quad (4)$$

The actual monthly mean maximum temperature data and that predicted by the above equation are shown in Fig. 2(a). The correlation coefficients between the actual and predicted data during training and validation periods are 0.9159 and 0.9211, respectively. The errors in the prediction during the training and the validation periods are shown in Fig. 2(b). The performance of genetic programming in prediction of monthly mean maximum temperature is summarized in Table 1.

To compare the performance of GP, same time series data is predicted by using ANN. A network topology of 5:5:1, learning rate of 0.5 and a momentum parameter of 0.9 is chosen for ANN by using trial and error approach. The best-trained neural network takes 100 epochs (113 sec) to give the lowest training error (RMSE = 1.3954). The actual monthly mean maximum temperature data alone with the ANN predicted data are shown in Fig. 3(a). The correlation coefficients between the actual and predicted data during training and validation periods are 0.9251 and 0.9278 respectively. The errors in the prediction during the learning and validation are shown in Fig. 3(b). The results obtained by using ANN are given in Table 1.

Table 1: Performance of GP and ANN in prediction of monthly mean maximum temperature

	GP		ANN	
	Training	Validation	Training	Validation
Average error	1.1149	1.1245	1.0819	1.0973
Root mean square error	1.4838	1.5079	1.3954	1.3996
Chi square error	0.0715	0.0721	0.0613	0.0661
Correlation coefficient	0.9159	0.9211	0.9251	0.9278
Number of epochs	123		100	
Time (sec)	165		113	

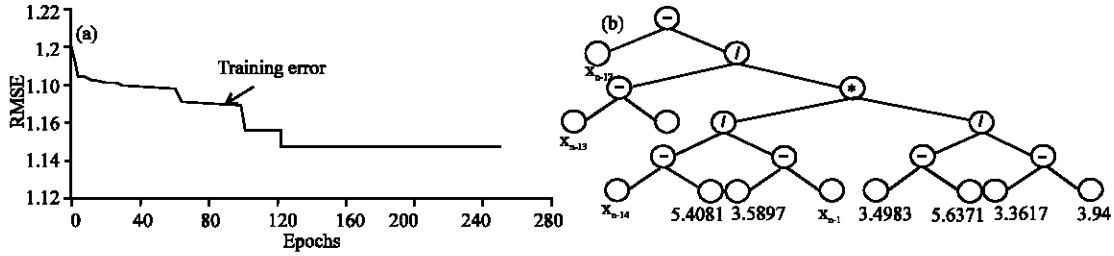

Fig. 1: (a) Error curve showing the convergence of GP during training with monthly mean maximum temperature data; (b) tree structure obtained by GP for monthly mean maximum temperature

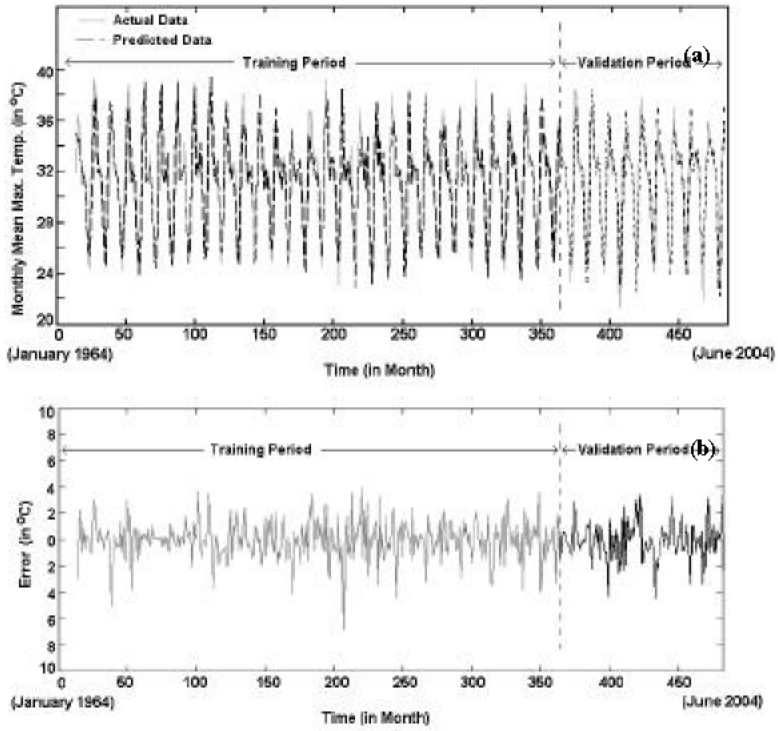

Fig. 2a: Actual and GP predicted monthly mean maximum temperature; (b) errors in the GP prediction during learning and validation of monthly mean maximum temperature

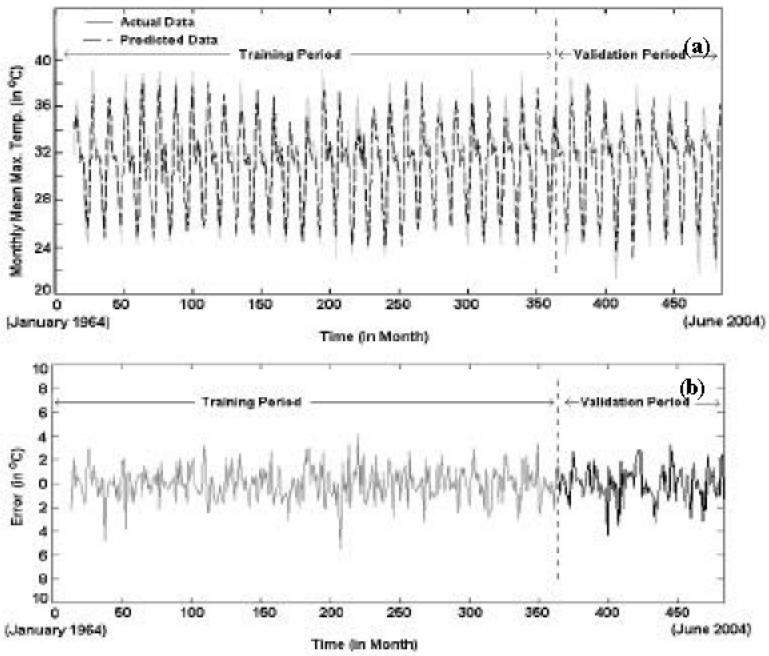

Fig. 3a: Actual and ANN predicted monthly mean maximum temperature; (b) errors in the ANN prediction during learning and validation of monthly mean maximum temperature

The study shows that model produced by GP can be used for the forecasting of monthly mean maximum temperature of Rajshahi city with a reasonable average error (±1.1245°C) in the prediction. On the other hand, ANN can be used for the prediction of same weather parameter with an average error of ±1.0973°C. In the present GP based modeling, only six function symbols viz. +, -, *, /, % and log and maximum depth of tree structure of seven are used. The result of GP could probably be enhanced by increasing number of function symbols, size of tree structure and more fine-tuning of GP parameters.

CONCLUSION

Genetic programming has been used in this paper for the forecasting of monthly mean maximum temperature of Rajshahi city. From the values of prediction errors and correlation coefficients between the actual and GP predicted climatic data, it is obvious that GP could be an efficient tool for the prediction of weather. Though the errors in the prediction produced by GP are high compared to ANN, the essence of GP is that it gives a model, which could be used, even by a layman, to predict weather manually. The future research can be carried to enhance the prediction capability of GP by increasing the number function symbols, tree-size and fine-tuning of GP parameters. As the conventional weather prediction involves tremendous amount of imprecision and uncertainty, GP could be an ideal tool for prediction.

REFERENCES

Angline, P.J., 1998. Evolving Predictors for Chaotic Time Series, Applications and Science of Computational Intelligence: Proceedings of SPIE, Volume 3390, Rogets *et al.* (Eds.), SPIE, Bellingham, WA, pp: 170-180.

Cortez, P., M. Rocha and J. Neves, 2001. Evolving Time Series Forecasting Neural Network Models. In Proc. of Int. Symposium on Adaptive Systems: Evolutionary Computation and Probabilistic Graphical Models (ISAS).

Coulibaly, P., 2004. Downscaling daily extreme temperatures with genetic programming. Geophysical Res. Lett., 31: L16203, 1-4.

Iba, H., 1995. A Numerical Approach to Genetic Programming for System Identification. Electrotechnical Laboratory Technical Report ETL.

Kishtawal, C.M., S. Basu, F. Patadia and P. K. Thapliyal, 2003. Forecasting summer rainfall over India using genetic algorithm, Geophysical Res. Lett., 30(23), 2203, 10.1029/2003GL018504.

Koza, J.R., 1992. Genetic Programming: On the Programming of Computers by Means of Natural Selection, MIT Press.

Maqsood, I., M.R. Khan and A. Abraham, 2002a. Intelligent weather monitoring systems using connectionist models, Int. J. Neural, Parallel and Sci. Computations, 10: 157-178.

Maqsood, I., M.R. Khan and A. Abraham, 2002b. Neuro-computing based Canadian weather analysis, The 2nd International Workshop on Intelligent Systems Design and Applications, Atlanta, Georgia, pp: 39-44.

Mondal, R.U. and S. Shahid, 2003. A Neural Network Approach for the Prediction of Monthly Mean Temperature, Islamic University J. Sci., 2: 7-12.

Oakley, H., 1994. Two Scientific Applications of Genetic Programming: Stack Filters and Non-Linear Equation Fitting to Chaotic Data. In Kinnear, Kim (Ed.), Advances in Genetic Programming. Cambrige, MA: The MIT Press, pp: 369-389.

Philip, N.S. and K.B. Joseph, 2001. On the Predictability of Rainfall in Kerala: An Application of ABF Neural Network, Proceedings of Workshop on Intelligent Systems Design and Applications (ISDA, 2001), In Conjunction with International Conference on Computational Sciences, ICCS, San Francisco.

Rashid, A., 1979. Geography of Bangladesh, Universal Press Ltd., New Delhi.

Zhang W., Z.M. Wu, G.K. Yang, 2004. Genetic programming-based chaotic time series modeling. J. Zhejiang Uni. Sci., 5: 1432-1439.