

Multi-Gene Genetic Programming for predict rainfall data

Mohammed Alweshah, Mohammed Ababneh and ¹Almahdi Alshareef

Al-Balqa Applied University, Jordan

¹Sabha University Sabha, Libyan Arab Jamahiriya

Abstract

Since the beginning of humanity, people have been interested in weather and environment such rainfall, humidity and temperature. A lot of environment specialist has tried to find a lot of methods that would allow them to know the weather in advance. This paper addresses the task of prediction rainfall data by depending on historical data and applying genetic algorithm. The quality of the data is the most important factor influence the performance of the genetic algorithm. The goal of this work is to predict the future rainfall amounts of the selected city as these predictions are very helpful and because rainfall prediction is very important for a lot of sectors such commercial, industrial, tourism and academic purposes. In this work, the genetic algorithm model is suitable for such purposes of predictions.

Keywords: Optimization, rainfall data, classification problem, genetic programming model.

1. Introduction

Predictive data mining is a search for patterns in data that can be represented or generalized to accurate future decisions (predictions) [1, 2]. In other way, it allows submitting records with some unknown field values and the algorithm system will predict the unknown output values based on previous patterns discovered from the historical data. The major tasks of predictive data mining include classification and regression. In the following sub-sections, classification and regression will be described.

Data classification is a common data mining task that deals with methods for assigning a set of input objects to a set of decision classes [3, 4]. The efficiency of the decisions is affected by the

performance of the classification task. In machine learning, the classification task is commonly referred to as supervised learning. In supervised learning, there is a specified set of classes, and example objects are labeled with the appropriate class. Numerous techniques have been used for rainfall predictions problems. Mislana and Haviluddin [5] use the Back Propagation Neural Network (BPNN).

Nayak, D. R., A. Mahapatra, et al [6] reports a detailed survey on rainfall predictions using different neural networks architectures over twenty five years. Also, a Fuzzy Set-Genetic Algorithms Model by Thongwan, T., A. Kangrang, et al [7] for estimating the rainfall. Cramer, S., M. Kampouridis, et al. (2017) [8]

using Genetic Programming for predicting rainfall in the context of rainfall derivatives [9-13].

In this work, the Multi Gene Genetic Programming (MGGP) will be used [14]. This paper aims to predict the rainfall for the upcoming years by using historical data regarding Darwin theory of evolution and natural selection. Moreover, the in-use algorithm should be checked in terms of accuracy and performance. Moreover; advantages of genetic algorithms will be explained. In addition, the software tools which be used in this work will be briefly discussed.

2. Material and methods

Genetic algorithms (GA), pioneered by Holland [15], are evolutionary search techniques inspired by natural selection (i.e survival of the fittest). The power of GA is being demonstrated for an increasing range of applications; financial imaging and future prediction [16]. The programs are evaluated against fitness function and the best solution selected for modification and re-evaluation. The modification-evaluation cycle is repeated until a correct program is resulted. Genetic programming (GP) is considered as specialization of genetic algorithm (GA) for evolution of executable programs [16].

The idea of getting new generations with a lot of parent's characteristics was modified to make many predictions and assumptions for more tasks. Reading future is a dream of everybody. It is very interesting to get predictions for some tasks such

the rainfall of upcoming years, temperature and others [15],.

The basics procedure for GA can be mentioned as follows:

1. Creating an initial population of random individuals. (The study population)
2. Testing the "fitness" of everyone.
3. If an individual is sufficiently good then the program will be stopped and the results successfully obtained and it will be ready to satisfy the predictor requirements.
4. Otherwise a new population (generation) will be created from the more fit individuals, using "genetic operators" such as reproduction, mutation, and crossover. These operators will be discussed in the upcoming sections.
5. Replacing the old population with the new population and returning to Step 2.

The aim of GP is to predict an accurate result through evolution-evaluation-selection cycle repetitions until a stopping criterion is met. The algorithm can be stopped when the maximum number of iterations reached or when the accepted error rate being achieved. This is occurred when the best fitness value being achieved.

3. Rainfall Prediction

Weather prediction or forecasting has been one of the most technologically and scientifically

challenging problems around the world in the last century [8]. The weather forecasts are divided into the following categories:

- Now casting; in which the details about the current weather or up to few hours are given.
- Short range forecasts (1 to 3 days); in which the weather, mainly rainfall, in each successive 24 hour intervals may be predicted up to 3 days.
- Medium range forecasts (4 to 10 days); average weather conditions and the weather on each day may be predicted with progressively lesser details and accuracy than that for short range forecasts.
- Long range forecasts (more than 10 days); there is no solid definition for long range forecasting which may range from monthly to a seasonal or annual forecast.

There are two approaches to predict rainfall; empirical and dynamical methods. The empirical approach is based on analysis of previous historical data of weather and its relationship to variety of atmospheric variables. The most widely empirical approaches used in climate prediction are regression, artificial neural networks, fuzzy logic, genetic algorithm and group method.

4. Proposed Method: Using Multi Gene Genetic Programming in Rainfall Prediction

In this work, MATLAB environment will be used for all project activities including preprocessing, post processing, classifications and performance evaluation. Genetic programming and symbolic regression (GPTIPS) used for rainfall data classification. Genetic programming (GP) based MATLAB is a very helpful tool to apply the spirit of evolution theory to predict the future features of a certain problem thus to predict the upcoming generations.

The algorithm toolbox is a collection of routines, which are written mostly in M-files which implement the most important functions in genetic algorithms. Figure 1 shows the important functions and work technique for GA.

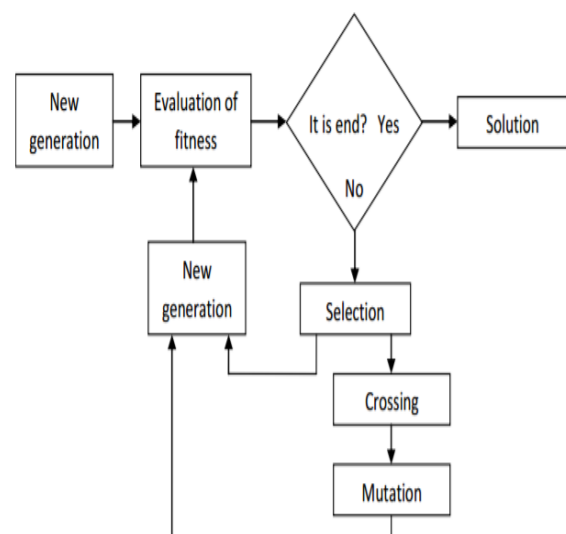


Fig 1 Functions of GA

4.1 Rainfall Data set used

In this section, the work based on data will be discussed. The data collected from Institute of Climate Change, University Kebangsaan Malaysia, covers the period from January -1978 till may- 2015 with a reading of rainfall every day. To accurately deal with such data, MATLAB-GPTIPS need to perform preprocessing by classification of the rainfall data to the ranges of heavy rain (31-60mm/day), moderate rain (11-30mm/day), light rain (3-10mm/day) and no rain (0-2mm/day). The historical data will be divided into two periods; training period and testing period.

Figure 4.1 is a sample of the picked rainfall data

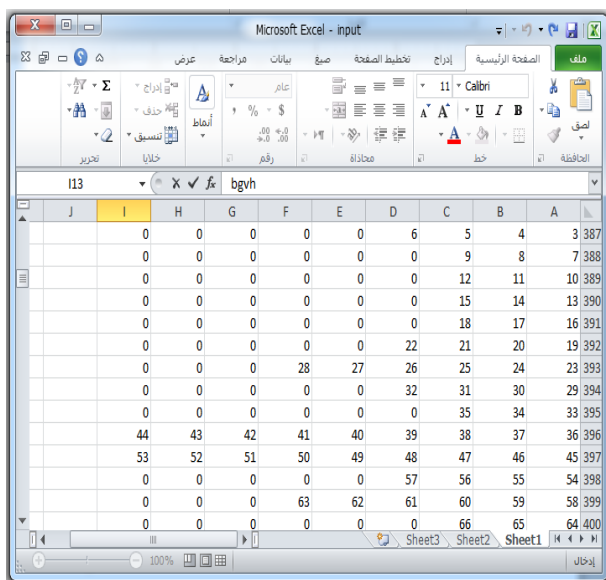


Figure 4.1 Rainfall Data Sample

4.2 Experimental Results

In this work, MATLAB –Multi Gene- Genetic Programing (MGGP) will be used in all phases including classification, pre-processing, off strings creating, result visualization and predictions.

To start building the model, different parameters have to be determined. These parameters include the population size, selection technique, crossover and mutation probabilities and the maximum number of genes allowed to constitute the multi-gene.

Table 5.1 Parameters setting

Parameter	Value
Population Size	200
Number of Generations	500
Selection Mechanism	Tournament
Tournament size	20
Maximum Tree Depth	4
Crossover Probability	.85
Mutation Probability	.1
Number of inputs	9
Maximum Gene	8
Function set	Times, minus, plus, sin
Constants range	[-10 10]

To judge the model, two performance evaluators were being used. Taking into consideration that the error is defined as the difference between predicted and actual outputs.

1- Root Mean Square of error (RMSE).

2- Regression (R^2)

In this work, a model structure to estimate the rainfall will be proposed through the use of multi-gene Genetic Programming (MG-GP). The data will be divided into two periods; training and testing to accurately walkthrough the project by ensuring acceptable performance rate.

To judge the performance of the algorithm (the model) , regression (R^2) has to be found. GPTIPS calculates $1 - R^2$ which is better to be close to zero. Figure 5.13 shows the $1 - R^2$ reference to the complexity of the model (Visually browsing a multigene regression model population).

Green dots represent the Pareto front of models in terms of model performance ($1 - R^2$) and model complexity. Blue dots represent non-Pareto models. The red circled dot represents the best model in the population as the ($1 - R^2$) is very close to zero. Clicking on a dot shows a yellow popup containing the model ID and the simplified model equation. Generated using the popbrowser function. This visualization may be used with the training, validation or testing data sets.

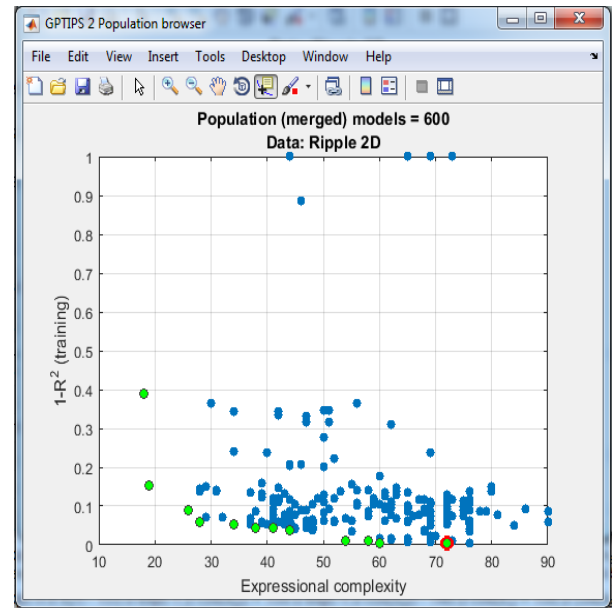


Figure 5.13 GPTIPS 2 Population browser

Figure 5.14 shows the close to linear relationship between the predicted and actual outputs. This is a common design pattern across a large number of GPTIPS functions. Performance scatterplots on training and testing data sets for a selected multigene regression model. Generated by the run tree function.

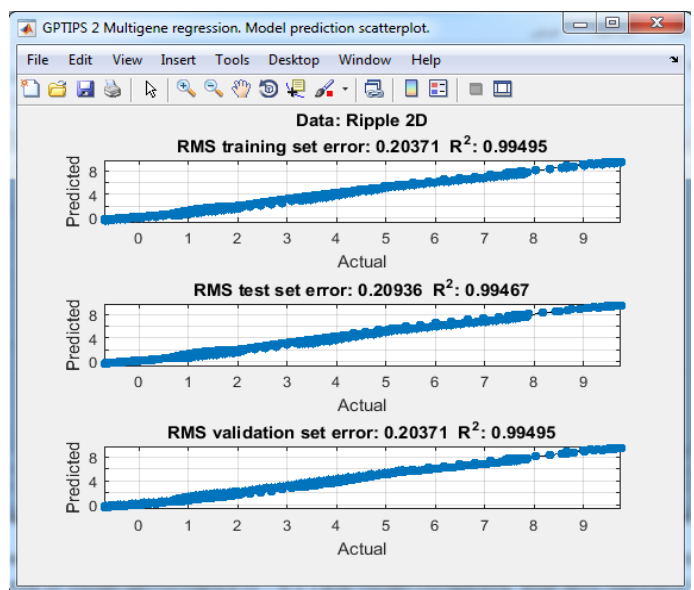


Figure 5.14 GPTIPS 2 Multigene regression, Model prediction scatterplot

The following figure 5.15 shows the multigene regression model predictions. As shown the performance functions of the model are the root mean square (RMS) of the error and the regression (R^2), the results are 0.20371 and 0.99495 respectively. The three rectangular charts in figure 5.4 represent the training, testing and validation.

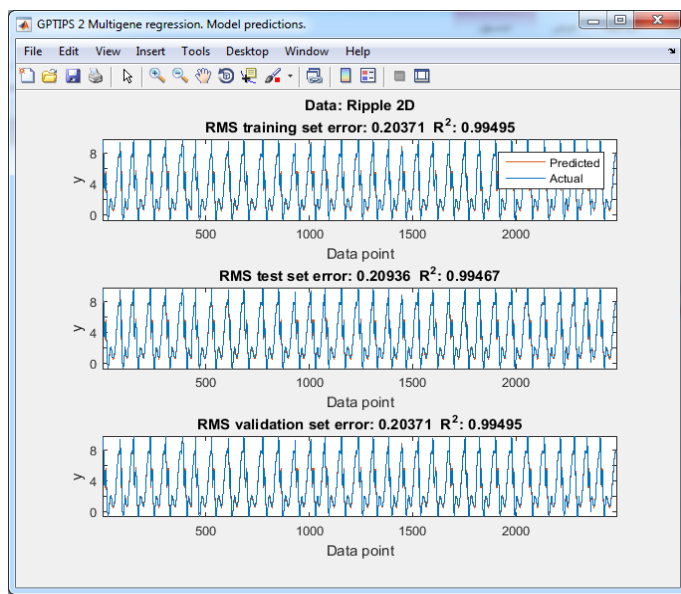


Figure 5.15 GPTIPS Multigene regression , Model predictions

Figure 5.16 shows the P-value of model genes (gene weights). It seems that there are some variance between the weights of the genes. Some of them are close to zero while others are positive and others are negative.

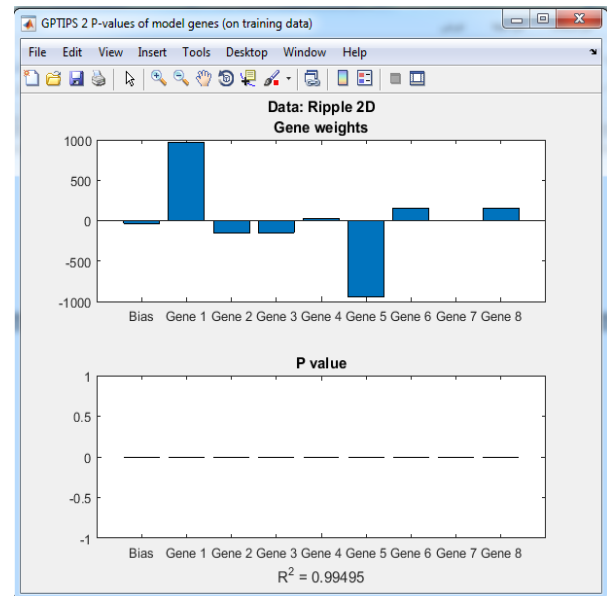


Figure 5.16 GPTIPS 2 P-values of model gene

5. Conclusion

Rainfall prediction is one of the most challenging tasks. The accurate rainfall prediction is helpful and useful for many sectors as already discussed. In this work, the genetic programming is being used to predict the rainfall for the upcoming years. The success of creating an accurate method of rainfall prediction will be an excellent work. Thus, testing and a lot of iterations is essential to achieve this goal. Clearly seen that this model is success with an error (RMSE) close to 0.2 and regression close to 0.99. This results clearly shows that genetic algorithm is suitable for such type of predictions.

6. Reference

- [1] I. H. Witten, E. Frank, M. A. Hall, and C. J. Pal, *Data Mining: Practical machine*

- learning tools and techniques*: Morgan Kaufmann, 2016.
- [2] A. M. Alshareef, A. A. Bakar, A. R. Hamdan, S. M. S. Abdullah, and M. Alweshah, "A case-based reasoning approach for pattern detection in Malaysia rainfall data," *International Journal of Big Data Intelligence*, vol. 2, pp. 285-302, 2015.
- [3] X. Wu, V. Kumar, J. R. Quinlan, J. Ghosh, Q. Yang, H. Motoda, G. J. McLachlan, A. Ng, B. Liu, and S. Y. Philip, "Top 10 algorithms in data mining," *Knowledge and information systems*, vol. 14, pp. 1-37, 2008.
- [4] M. Alweshah, A. Omar, J. Alzubi, and S. Alaqeel, "Solving Attribute Reduction Problem using Wrapper Genetic Programming," *International Journal of Computer Science and Network Security*, vol. 16, pp. 77- 84, 2016.
- [5] S. Hardwinarto and M. Aipassa, "Rainfall Monthly Prediction Based on Artificial Neural Network: A Case Study in Tenggara Station, East Kalimantan-Indonesia," *Procedia Computer Science*, vol. 59, pp. 142-151, 2015.
- [6] D. R. Nayak, A. Mahapatra, and P. Mishra, "A survey on rainfall prediction using artificial neural network," *International Journal Of Computer Applications*, vol. 72, 2013.
- [7] T. Thongwan, A. Kangrang, and S. Homwuttiwong, "An Estimation of Rainfall using Fuzzy Set-Genetic Algorithms Model," *American Journal of Engineering and Applied Sciences*, vol. 4, 2011.
- [8] S. Cramer, M. Kampouridis, A. A. Freitas, and A. K. Alexandridis, "An extensive evaluation of seven machine learning methods for rainfall prediction in weather derivatives," *Expert systems with applications*, 2017.
- [9] A. A. Afifi and S. P. Azen, *Statistical Analysis: A Computer Oriented Approach*. Orlando, FL, USA: Academic Press, Inc, 1979.
- [10] M. Alweshah and S. Abdullah, "Hybridizing firefly algorithms with a probabilistic neural network for solving classification problems," *Applied Soft Computing*, vol. 35, pp. 513-524, 2015.
- [11] M. Alweshah, A. I. Hammouri, H. Rashaideh, M. Ababneh, and H. Tayyeb, "Solving Time Series Classification Problems Using Combined of Support Vector Machine and Neural Network," *International Journal of Data Analysis Techniques and Strategies* vol. In press, 2017 2017.
- [13] M. Alweshah, A. I. Hammouri, and S. Tedmori, "Biogeography-based optimisation for data classification problems," *International Journal of Data Mining, Modelling and Management*, vol. 9, pp. 142-162, 2017.
- [14] D. P. Searson, D. E. Leahy, and M. J. Willis, "GPTIPS: an open source genetic programming toolbox for multigene symbolic regression," in *Proceedings of the International multiconference of engineers and computer scientists*, 2010, pp. 77-80.
- [15] L. B. Booker, D. E. Goldberg, and J. H. Holland, "Classifier systems and genetic algorithms," *Artificial Intelligence*, vol. 40, pp. 235-282, 1989.
- [16] J. K. E. Kinnear, "Derivative methods in genetic programming," *Evolutionary Computation*, vol. 1, pp. 103-113, 2000.