M5 Model Tree and Gene Expression Programming for the Prediction of Metrological Parameters

CONFERENCE PAPER · NOVEMBER 2015

DOI: 10.13140/RG.2.1.3315.2085

1 AUTHOR:

Khalid Raza

Jamia Millia Islamia

21 PUBLICATIONS **50** CITATIONS

SEE PROFILE

Proceeding of *IEEE 2015 International Conference on Computers, Communications, and Systems (ICCCS-2015)*, Nov 2-3, 2015, Kanyakumari, India

M5 Model Tree and Gene Expression Programming for the Prediction of Metrological Parameters

Khalid Raza
Department of Computer Science,
Jamia Millia Islamia (Central University)
Jamia Nagar, New Delhi-110025
kraza@jmi.ac.in

***Abstract*— Metrological parameter prediction is an important task carried out by metrological researchers and water resource analyst for the estimation of water demand, water resource planning, management and operation. Accurate predictions of these parameters are still an open problem. M5 model tree and gene expression programming are two popular machine learning techniques which are considered as white-box model. In this paper, we have applied two machine learning techniques, such as M5 model tree and gene expression programming, for the prediction of various metrological parameters at a weather station. Among these two techniques, one is better over other for the prediction of a set of parameters.**

Keywords- Metrological variable predictions; Gene expression programming; Model tree; Evolutionary algorithm

I. INTRODUCTION

Metrological parameters, such as temperature, evaporation, humidity, etc., plays important role in many areas of science and engineering including metrology, agriculture, water resource and disaster management. Hence, its prediction is an important task which is still an open problem for the researchers. Numeric Prediction is the task of predicting continuous values for given set of inputs. These problems may be solved by linear regression methods, and can also be tackled by deploying transformation to the variables so that nonlinear problems are transformed to a linear one. M5 model tree and gene expression programming are two popular machine learning techniques which are widely used for numerical prediction of continuous values. The motivation of using these techniques is that both are white-box modeling tool. Unlike artificial neural network (ANN), internal learnt parameters of both the techniques can be extracted. Both the techniques learn from the data, and M5 model tree represents the prediction equations in the form of a number of rules, while gene expression programming (GEP) store it in the form of trees.

Guven and Aytek [1] applied gene expression programming for finding stage–discharge relationship. The time series of daily stage and discharge were used from two metrological stations, Schuylkill River at Berne, and Schuylkill River at Philadelphia, collected from US Geological Survey. The discharge (Q) is modeled in terms of the stage (h) using the GEP approach. After training and testing, root mean square error (RMSE) and determination coefficient (R_2) for each model have been computed that shows that stage having zero lag [h(t)] corresponding to highest RMSE and R_2 for both downstream as well as upstream stations. It has been observed that adding Q with one-day lag [Q(t–1)] to single h(t) has worsened the GEP's performance. Also, addition of Q(t–2) to [h(t), Q(t–1)] contributes with higher RMSE and R_2 for both the stations. Barbulescu and Bautu [2] applied an adaptive gene expression programming (AdaGEP) for meteorological time series modeling. AdaGEP identifies the appropriate number of genes and applied an adaptive gene deactivation mechanism automatically. Each chromosome of AdaGEP is enhanced with a bit string called "gene map" whose size equal to the number of genes in GEP individuals. The search strategy used by GEP let it identify the variables automatically that are playing major role in estimating future values based on past 'n' inputs. The window size (w) of 5 performed well over all other runs.

The authors in [3] applied M5 for estimating reference Evapotranspiration (ET.) using Pan Evaporation. They developed pan coefficient (Kp) equation using indicator regression. The Kp values has been taken as function of daily mean relative humidity (RH %), and daily mean wind speed. Onyari and Ilunga [4] applied multilayer neural network and M5 model tree for the prediction of stream-flow of Luvuvhu catchment located in the north-eastern part of South Africa. Sattari et al. [5] developed monthly model for evapotranspiration of an area around Ankara, Turkey. Meteorological data for the time period between 1975–2006 were considered. Different combinations of metrological parameters have been used as input for the prediction purpose. The M5 model predicts evapotranspiration with highest correlation coefficient of 0.997 and MSE of 0.002. In [6], authors compared M5 model tree with ANN in their prediction accuracy. Here, M5 model tree estimated evapotranspiration with a correlation coefficient of 0.98 and a mean absolute error of 0.270, but ANN has correlation coefficient and mean absolute error were 0.975 and 0.33, respectively. Hence, work of [6] shows betterment of M5 over ANN. In [7], three different types of neural networks viz. feedforward neural network (FFNN), radial basis function (RBF), recurrent neural networks (RNN) have been deployed for the prediction of seven metrological parameters of a weather station. Among these three different neural networks, RNN is found to perform well. The highest correlation coefficient of maximum temperature, minimum temperature and humidity determined to be 0.92, 0.85 and 0.82, respectively.

The paper is organized as follows. Section II describes the methodologies and datasets used in this paper. Section III presents the results and discussions and finally Section IV concludes the paper.

II. MATERIALS AND METHODS

In this paper, we applied M5 model tree and Gene Expression Programming (GEP) for numerical prediction of continuous values. The 17 years daily data of New Delhi/Palam metrological station has been used for the period 1997 – 2013 for training and testing. Seven different metrological variables are considered that includes maximum temperature, minimum temperature, average temperature, humidity, visibility, wind speed, and maximum sustained wind speed.

A. M5 Model Tree

Model Tree is a tree based machine learning technique which deals with numeric continuous values and its learning algorithm is called M5, proposed by Quinlan in 1992 [9]. The main advantage of this machine learning approach is that it is like a white-box learning model that gives us a set of mathematical expressions that shows dependencies between variables. M5 model tree algorithm was further modified in M5flex (for dealing with multidimensional data), and M5opt (greedy and non-greedy approach is combined together). M5 model tree is easy to use and it is robust when it comes in finding missing data. Some of the basic advantages of model tree are: i) it performs variable screening or feature selection implicitly, ii) needs relatively less effort for data preparation, iii) nonlinear relationships among variables do not affect the performance of tree, and iv) tree representations are better for analytics and easy to interpret. Decision trees are generated for regression problems using splitting and pruning. Model tree is built and the best leaf is taken as a rule. The dependent variable is predicted by the use of finite number of unordered values.

Model tree is based on two concepts i.e. conventional decision tree and linear regression functions. Two stages used by M5 model tree are splitting and pruning. Two steps undertaken for building model trees are:

First step: Decision tree induction algorithm is applied to construct the tree. Splitting criterion of decision tree minimizes the intra subset variation. The splitting criterion calculates the expected reduction in error by treating the standard deviation (SD) of the values as the measure of error at the nodes. Splitting in M5 stops when instances either vary very marginally or a few instances left.

Second step: Pruning of trees to interior nodes is done and the nodes are replaced by regression plane instead of constant value. Expected error at each node is estimated by pruning. Firstly, absolute difference between the actual and the predicted value is averaged for each node.

a) Splitting: In splitting criterion a set T is taken and its input is split into subsets T1, T2,.., Tn. These subsets are further split into simpler subsets. This splitting process continues until no more instances are left to split or when output value of the instances remained vary very slightly. M5 follows greedy approach so to minimize error at each nodes, Standard Deviation Reduction (SDR) is computed by taking one node at a time, which is given by,

$$SDR = \frac{SD(T) - \sum SD(T_i)|T_i|}{|T|} \quad (1)$$

Where,

T = set of examples that reach the node; T_1, T_2, ..T_n = sets that result from splitting the node according to the chosen attribute, and SD = standard deviation.

b) Pruning: It is a technique used to remove those sections of the tree which may not be useful in later stages. It reduces the complexity of the classifier, reduces the erroneous and noisy data, and improves the results by better prediction of the data. Using standard regression errors are calculated at each node and according to the error terms are dropped. The difference between the actual and the expected values are calculated at each node. There are two types of pruning: i) reduced error pruning, where the most popular class replace the nodes and starting at the leaves. It simplifies the data thus increasing the speed. ii) Cost complexity pruning (CCP), which is computed as,

$$CCP = \frac{err(prune(T,t),S) - err(T,S)}{|leaves(T)| - |leaves(prune(T,t))|} \quad (2)$$

Where, err(T, S) = error rate of tree T over data set S, Prune(T,t) = tree is obtained by pruning the subtrees t from the tree T.

c) Smoothing: Sharp discontinuities of the subtree are avoided by smoothing and the value combined with the predicted value of the linear model of that node. It increases the accuracy of prediction by smoothing the sharp nodes of the adjacent models.

B. Gene Expression Programming

Gene Expression Programming (GEP) is the learning algorithm which learns from the data and find out the relationships between variables. It also constructs models that explain these relationships. It was proposed by Ferreira in 2001 [8]. The GEP is an evolutionary algorithm and related to genetic algorithm (GA) and genetic programming (GP). It adopts expressive parse tree of different shape and size from GP and linear chromosomes of fixed length from GA. The parse trees act as phenotype, creating a genotype/phenotype system, while linear chromosomes act as genotype. Hence, by this way GEP is multi-genetic which encodes multiple parse trees in every chromosome.

The GEP generates population of individuals that represents models or solutions, performs selection and reproduction based on some fitness function, and applies genetic operators such as mutation, cross-over or recombination to introduce genetic variations. GEP can solve many problems that cannot be solved by GA and GP. It is also much faster than GA and GP on solving particular

problems. It can be applied in regression, classification, neural network etc. The main difference among GA, GP and GEP resides in the nature of its individuals. In GAs, individuals are linear strings of fixed length, in GP the individuals are nonlinear entities of fixed shapes and size, and in GEP the individuals are encoded in the form of linear strings of fixed length that further represented as non-linear entities of different shapes and size (e.g. expression trees).

a) Description of GEP Algorithm

The process starts with the random generation of chromosomes as initial population, then the chromosomes are expressed and the fitness of each individual are computed. The individuals are then selected based on some fitness and reproduced. The individuals of new generation are, in their turn, subjected to the same developmental process: expression of the genomes, confrontation of the selection environment, and reproduction with modification. These process is repeated for a certain number of generations or until algorithm has reached to the solution [8]. Here, reproduction includes both replication and genetic operator creating genetic diversity. The genetic operators randomly pick the chromosomes that need to be modified. Hence, a chromosome may be modified by either one of several operators at a time, or not modified at all [8].

In GEP, chromosomes comprises of a linear, symbolic string of fixed length, composed of one or more genes. Despite their fixed length, GEP chromosomes is able to code expression three having different shapes and sizes.

b) Fitness Functions

In regression problem, the dependent variable is continuous values and therefore, the output of the model must also be continuous. Thus, evaluating the fitness of evolving models is quite straightforward. It can be done by comparing the output of the model to the actual value in the training dataset. Some of the mostly used functions are mean squared error (MSE), root mean square error (RMSE), mean absolute error (MAE), and so on. For regression problem, fitness functions which are based on Pearson's correlation coefficient are very smooth and comparable with other measures.

C. Datasets

Metrological data have been compiled from [12] for the period 1997 to 2013 of New Delhi/Palam weather station-421810 which is situated at a latitude of 28.56, longitude of 77.11 and at an altitude of 220. Daily observed metrological parameters are maximum temperature, minimum temperature, average temperature, humidity, visibility, wind speed, and maximum sustained wind speed. We have considered these seven metrological variables as input dataset to predict the one of these seven variables at a time. When applying regressions on time-series data, it is necessary to include lagged values of the dependent variable as independent variables. Here, we lagged all the dataset by one day so that based on today's data we can predict tomorrow's values.

III. Results and Discussions

We applied both M5 model tree and GEP separately for the prediction of seven metrological variables. For the simulation study of M5 model tree, Weka software tool [10] has been used. For GEP, GeneXproTool 5.0 [11] has been applied. GeneXproTool 5.0 is an extremely flexible data modeling tool based on GEP and developed for regression and time series prediction modeling. GeneXproTools supports tools for data cleaning, data analysis, models generation, integration of generated models with other applications, and translate the code to up to 19 different programming languages. The daily observed data of different variables of New Delhi/Palam metrological station has been taken for the period 1997-2013 (17 years) for training and testing. Out of these datasets, 70% data has been used for training and remaining 30% used for testing purpose. The datasets includes maximum temperature, minimum temperature, average temperature, humidity, visibility, wind speed, and maximum sustained wind speed. The dataset has been lagged by one-day to predict a meteorological parameter of next day by using today's data. Since we have considered seven meteorological parameter, hence seven different models has been developed. Each model takes seven inputs and predicts one parameter at a time. The predicted results of M5 model tree are shown in Table 1.

Out of seven parameters, M5 model tree is able to predict all three variables related to temperature with correlation coefficient above 0.95. Other three parameters such as humidity, wind speed and maximum sustained wind speed have correlation coefficient of 0.84, 0.82 and 0.81, respectively. The visibility has the lowest correlation coefficient of 0.70 among the seven parameters. The last column of the Table 1 shows number of rules identified by M5 model tree. The highest number of rules is 45 for humidity, while lowest number of rules is for maximum sustained wind speed.

Gene expression programming has been applied on the same dataset for the prediction of all the seven parameters. Seven different GEP prediction models have been developed, each one taking seven inputs to predict single parameter at a time. The prediction accuracy of GEP model is presented in Table 2. Among the seven variables, average temperature is predicted with highest correlation coefficient of 0.96. When the results of GEP are compared with M5 model tree, it is found that minimum temperature, humidity and visibility prediction of GEP is better over M5 model tree. Actual versus predicted graphs are for different parameters are shown in Fig. 1.

TABLE I. Predicted results of M5 Model Tree

Metrological Variables	Correlation Coefficient	MAE	Number of Rules
Max. Temperature (T_{max})	0.9801	1.0774	38
Min. Temperature (T_{min})	0.9453	1.5221	30
Avg. Temperature (T_{avg})	0.9936	0.6247	37
Humidity	0.8346	8.5657	45
Visibility	0.7016	0.5089	16
Wind Speed	0.8151	1.866	14
Max. Sustained Wind Speed	0.8123	3.3734	10

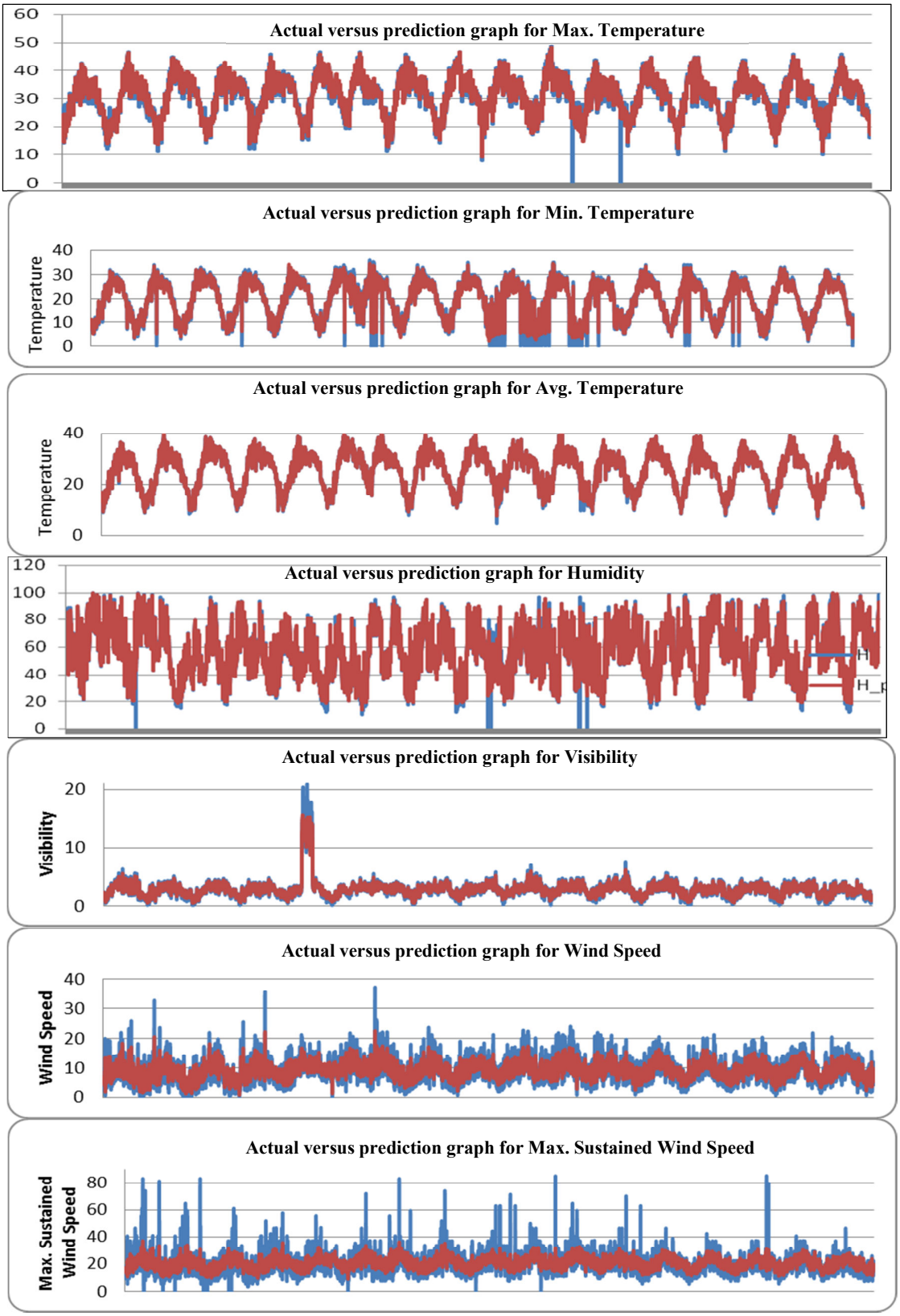

Figure 1. Actual versus predicted graph for all the seven metrological variables using GEP
(Blue: actual, Red: predicted)

TABLE II. PREDICTED RESULTS OF GENE EXPRESSION PROGRAMMING

Metrological Variables	Correlation Coefficient	MAE
Max. Temperature (T_{max})	0.9428	0.8887
Min. Temperature (T_{min})	0.9576	0.9170
Avg. Temperature (T_{avg})	0.9549	0.9119
Humidity	0.9132	0.8340
Visibility	0.8724	0.8200
Wind Speed	0.4908	0.5752
Max. Sustained Wind Speed	0.5741	0.3296

We compared our obtained results with a neural network based approach in [7] and found that both M5 model tree and gene expression programming works better for the prediction of maximum temperature, minimum temperature and humidity.

IV. CONCLUSIONS AND FUTURE DIRECTIONS

This paper presents an application of M5 Model Tree and Gene Expression Programming for the prediction of seven metrological parameters at a metrological station. The motivation for using M5 Model Tree and Gene Expression Programming over other techniques such as artificial neural network is that these machine learning techniques work like a white-box model. The M5 gives the prediction equations in the form of a number of rules, white Gene Expression Programming represents it in the form of expression trees. Among these two applied techniques, M5 Model Tree found to perform better over Gene Expression Programming for the prediction of maximum temperature, average temperature, wind speed and maximum sustained wind speed. On the other hand, GEP performs better for the prediction of minimum temperature, humidity and visibility. Among the seven metrological parameters, average temperature is predicted with highest accuracy having correlation coefficient of 0.99.

In this paper, we considered M5 Model Tree and Gene Expression Programming which are based on Multiple Input Single Output (MISO) model. So, for future work, we would apply the principle of Multiple Input Multiple Output (MIMO) model. This MIMO model will help us to have a single model for the prediction of all the parameters in a single go.

ACKNOWLEDGMENT

The author would like to thanks Ms. Sidrah, Mr. Kashif Shamshi and Mr. Suhail for compilation of metrological data from http://www.tutiempo.net.

REFERENCES

[1] A. Guven and A. Aytek, "New approach for stage–discharge relationship: gene-expression programming," Journal of Hydrologic Engineering, vol. 14, no. 8, 2009, pp. 812-820.

[2] A. Barbulescu and E. Bautu, "Meteorological time series modeling using an adaptive gene expression programming," Proc. 10th WSEAS international conference on evolutionary computing, 2009 (pp. 17-22). World Scientific and Engineering Academy and Society (WSEAS)

[3] P. Ditthakit and C. Chinnarasri, "Estimation of Pan Coefficient using M5 Model Tree", School of Engineering and Resources, Walailak University, Nakhon Si Thammarat 80160, Thailand, 2012.

[4] E. K. Onyari and F. M. Ilunga, "Application of MLP Neural Network and M5P Model Tree in Predicting Stream flow: A case study of Luvuvtu Catchment, South Africa," International Journal of Innovation, Management and Technology, vol. .4, no.1, Feb.2013.

[5] M.T. Sattari, M. Pal, K. Yurekli, and A. Ünlukara, "M5 model trees and neural networks based modelling of ET0 in Ankara, Turkey," Turkish Journal of Engineering and Environmental Sciences, vol. 37, issue 2, 2013, pp. 211-219..

[6] N. Ghahreman and M. Sameti, "Comparison of M5 model tree and Artificial Neural Network for estimating Potential Evapotranspiration in semi- arid climates," Department of Irrigation and Reclamation Engineering, University of Tehran, Karaj, Iran, March 2014.

[7] K. Raza and V. Jothiprakash, "Multi-output ANN model for prediction of seven meteorological parameters in a weather station," J. Inst. Eng. India Ser. A, vol. 95, issue 4, 2014, pp. 221-229, doi: 10.1007/s40030-014-0092-9

[8] C. Ferreira, "Gene expression programming: a new adaptive algorithm for solving problems," Complex Systems, vol. 13, issue 2, 2001, pp. 87-129.

[9] J.R. Quinlan, "Learning with continuous classes," Proc. 5th Australian joint conference on artificial intelligence (Vol. 92, pp. 343-348), 1992.

[10] M. Hall, E. Frankdrop, G. Holmes, B. Pfahringer, P. Reutemann, I.H. Witten, "The WEKA data mining software: an update," SIGKDD Explorations, Vol. 11, Issue 1, 2009.

[11] GeneXproTools 5.0 by GepSoft, http://www.gepsoft.com/

[12] http://www.tutiempo.net/ accessed on October 20, 2014.

Cite this article as:

Raza, K. (2015). **M5 Model Tree and Gene Expression Programming for the Prediction of Metrological Parameters.** In Proc. of *IEEE 2015 International Conference on Computers, Communications, and Systems (ICCCS-2015)*, Nov 2-3, 2015, Kanyakumari, India, p. 47-51, IEEE.