Ramsundram et al., Irrigat Drainage Sys Eng 2016, 5:3
DOI: 10.4172/2168-9768.1000175

Irrigation & Drainage Systems Engineering

Review Article Open Access

Comparison of Decision Tree Based Rainfall Prediction Model with Data Driven Model Considering Climatic Variables

Ramsundram N*, Sathya S and Karthikeyan S

Department of Civil Engineering, Kumaraguru College of Technology, Coimbatore, Tamilnadu, India

Abstract

In hydrological cycle, precipitation initiates the flow and governs the system. The preciseness in the prediction of rainfall will reduce the uncertainty involved in estimating the associated hydrological variables such as runoff, infiltration, and stream flow. Many research works has been channelled towards improving the accuracy of these predictions. ANN is the most widely used neural networks in Integrated Water Resource Management. Most of these models, utilize the strength of data-driven modelling approach. The reliability of these predictions depends on the preciseness in selecting the correlated variables. If the available historical database fails to record the most correlated variable, then reliability on these data-driven approach predictions is questionable. In this paper, an attempt has been made to develop a methodological framework that utilizes the strength of a predictive data-mining analysis (decision tree). The developed decision tree based rainfall prediction model maps the climatic variables, namely; a) temperature, b) humidity, and c) wind speed over the observed rainfall database. The performance of the developed model is evaluated based on three performance indicators (Nash Sutcliffe efficiency, RMSE and MSE). The performance of the developed model is also compared with the well- known data-driven (Artificial Neural Network) based rainfall prediction model.

Keywords: Rainfall prediction; Data-mining; Decision tree; Artificial neural network

Introduction

The decisions taken by the policy makers or research community on water resource management depend upon the predicted/forecasted scenarios generated by the models. The dependability on the end results of those models is based on accuracy at which the input parameters for the model are measured or predicted. Especially in the case of hydrological cycle is concerned, precipitation plays a vital role in initiating the entire process. The dependability on the predicted/forecasted precipitation will ensure the reliability in the estimation of the other dependable hydrological variables. To ensure the dependability on the precipitation predictions, in the recent past many models had been proposed namely time series prediction and data driven models. In the case of prediction or forecast based on previous rainfall events is concerned, the most popular and well known technique of time series forecast is regression analysis and moving average. To utilize the knowledge of regression and moving average in a combined nature, ARIMA model has been proposed. Further, to increase the level of prediction of ARIMA model many improvements had been suggested by researchers. In these statistical techniques, the present time rainfall depends upon the pervious rainfall and statistical parameters of the historical rainfall database. It conveys that statistical parameters derived from historical database govern the future rainfall events. The accuracy in statistical parameter is ensured by the length of historical database. If the database is of limited/sparse in number, then dependability on the time series forecasted rainfall pose lot of uncertainty.

On the other hand, to utilize the hidden relationship or correlation existing between rainfall events, data driven models namely artificial neural networks (ANN), genetic programming (GP), etc., are proposed by research community in recent past decade and still it is beginning an area of research. Similarly, in case of well reported data driven models such as ANN and GP, the knowledge derived from the training database ensure the accuracy in forecast. In both modelling approaches success history has be reported well with huge historical database. When historical database is of limited nature i.e., the length of database used for arriving at statistical parameters or for training are few in number, then it may result in under or over prediction. To improve the ANN from under or over prediction of pattern, many improvements had been reported. However, the trained ANN model fails to express the mapped relationship explicitly between the input and output variables. To overcome this above drawback of data-driven models, in this research the developed methodological framework utilizes the strength and capability of recently developed data-mining algorithms to recover the knowledge from large historical database. Data mining is a process which tries to identify information or make sense of large database mostly comprising of unsupervised data [1]. This process or technique utilizes the strength/capability of statistical analysis, database systems, machine learning and pattern recognition. The capability drawn from data-driven model will mislead researchers to think both data-mining and data-driven model are one other the same. As per the definition stated by Cios et al. it can be clearly understood that the data mining process looks for the sensible patterns hidden in the large database, where as data-driven models generates patterns based on the learning process [1].

Data mining found its development and application in the recent past for analyzing market based transaction, financial transaction etc. It developed knowledge discovery process (KDP) for application in engineering field. Very few research works are reported the application of data mining particularly to water resources [2,3]. Data-mining approach has been used to derive the hidden relationships between reservoir releases and other variables. Decision tree (DT) is one of the data-mining has been used to derive the reservoir operation rules and evaluated with rules derived from linear programming optimization model and found that the DT performs better than linear programming

***Corresponding author:** Ramsundram N, Department of Civil Engineering, Kumaraguru College of Technology, Coimbatore, Tamilnadu, India, Tel: 0422-2661100; E-mail: ramsundram.civil@kct.ac.in

Received December 02, 2016; **Accepted** December 07, 2016; **Published** December 14, 2016

Citation: Ramsundram N, Sathya S, Karthikeyan S (2016) Comparison of Decision Tree Based Rainfall Prediction Model with Data Driven Model Considering Climatic Variables. Irrigat Drainage Sys Eng 5: 175. doi: 10.4172/2168-9768.1000175

[2,3]. Further, the efficiency of DT algorithm is evaluated by comparing with the optimal rules derived from regression based rules for flood control reservoir (Wei and Hsu, 2008), and inferred that DT based reservoir rule curves performed better than the regression based rules. The rules derived from DT algorithm are easier to understand from explicit expression between the attributes and the decision-making variable. Data mining has been widely used in the area of reservoir operation for extracting the hidden relationship that exists in the historical database and utilizing the same for predicting the future reservoir operation rules. In this research work an attempt has been made to incorporate the strength of DT data-mining algorithm to recover the relationship that exists between climatic variables and the rainfall historical database. The developed framework utilizes the recovered knowledge for predicting the future rainfall events.

ANN is a complex data driven model to represent non-linear input/output relationship. ANN can be applied in various fields, viz. approximation, pattern recognition and classification, optimization and prediction [4]. ANN is based on training, not on statistical or analytical assumption. ANN model can be trained to predict results from information at very high speed [5]. The friction factor of the external flow over pile of circular tubes is estimated by using both multiple linear regression and ANN approaches. The better results were observed in ANN [6]. Azimian has also applied ANN to predict the friction factor for the flow inside a pipe [7]. The back propagation method, a gradient-descent algorithm that minimizes the error between the output of the training input/output pairs and the actual network outputs, is used to adjust the connecting weights [8]. It has justified that ANN predicts highly complex relationships between the input and the output variables. The research on rainfall modeling proposed. Prediction of daily inflow of reservoir.

Methodology

For achieving the desired objective of this proposed research, a methodological framework has been developed by utilizing the strength of descriptive data mining (Decision tree) algorithm for recovering hidden relationships that exists in the rainfall and climatic variables. The developed methodology (Figure 1) has three major process flows, namely; a) data pre-processing and data selection, b) extracting hidden relationships and developing decision tree, and c) rainfall prediction model.

Data preprocessing and data selection

Preprocessing is done to remove the noise that might be present in the database due to error in observation, and also to fill the missing values in the database. In case of climatic variables, the database might have record on wind speed, wind direction, maximum and minimum temperature, humidity, and sun light hours with respect to time period. For a year period the size of the database might be [360, 6]. Extracting information from this huge historical database might lead to poor knowledge recovery from the database [9]. In this study to improve the knowledge recovery process, as a part of data preprocessing important and well correlated reservoir variables were selected based on correlation a statistical technique.

Correlation

The historical database of rainfall and climatic parameters has been prepared and the same is subjected to correlation analysis considering rainfall as dependent variables and other climatic parameters as independent variables.

$$\text{Correlation coefficient} = \frac{\sum\left(x-\bar{x}\right)\left(y-\bar{y}\right)}{\left(n-1\right)s_x s_y}$$

where, S_x and S_y represent the sample standard deviations of the x and y data values, respectively.

Extracting hidden relationship

The developed framework utilizes the strength of decision tree (DT) predictive data-mining algorithm for extracting the hidden relationship that exists in the historical database.

Decision tree

A decision tree is a decision support tool that uses a tree-like graph or model of decisions and their possible consequences, including chance event outcomes, resource costs, and utility. It is one way to display an algorithm. Decision trees are commonly used in operations research, specifically indecision analysis, to help identify a strategy most likely to reach a goal [10]. Decision tree is very simple to understand and interpret. Decision tree is commonly used in data mining process. Decision tree is able to handle both numerical and categorical data and it also performs well with large databases. In decision tree analysis, it is possible to validate a model using statistical tests.

Rainfall prediction model

Figure 1 presents the developed Rainfall Prediction data-mining (RPD) Framework which explores the historical database for temporal relationships that exists between the reservoir releases and other reservoir variables [11]. The developed RPD framework process begins with identification of important climatic variables that are temporally correlated with rainfall.

To identify the temporal dependence, the RPD framework uses temporal correlation analysis. The identified important reservoir variables form the input for predicting the rainfall at time't'. The segregated database of important climatic variables that are correlated with rainfall then subjected to data preparation step or procedure. In the data preparation step, the segregated database has been analyzed for presence of noise or outliers. The prepared databases have been split into training and testing datasets. From the total length of the historical reservoir operation database, 80% of the datasets are grouped as training dataset and the remaining 20 % of the data length are termed as test datasets [12]. The prepared training dataset has been subjected to data-mining process. The chosen three data-mining algorithms mines or maps the input climatic variables with respective rainfall, in this process the data-mining algorithms estimate there modeling parameters number of nodes for DT. The recovered temporal knowledge has been considered as guide rules for classifying or predicting the future rainfall for the input variable combination that exist in the test dataset.

Artificial neural networking

A computational model based on number of simple, highly interconnected processing elements, which process information by their dynamic state of response to external inputs. ANN is a non-linear mathematical system derived from the concept of neuron function in the human brain. ANN's are considered non-linear statistical data modeling tools where the complex relationship between inputs and outputs are modeled or patterns are found [13]. It can actually learn from observing database. It is a random function approximation tool. Back propagation type neural network process information in interconnecting processing element termed nodes. The back

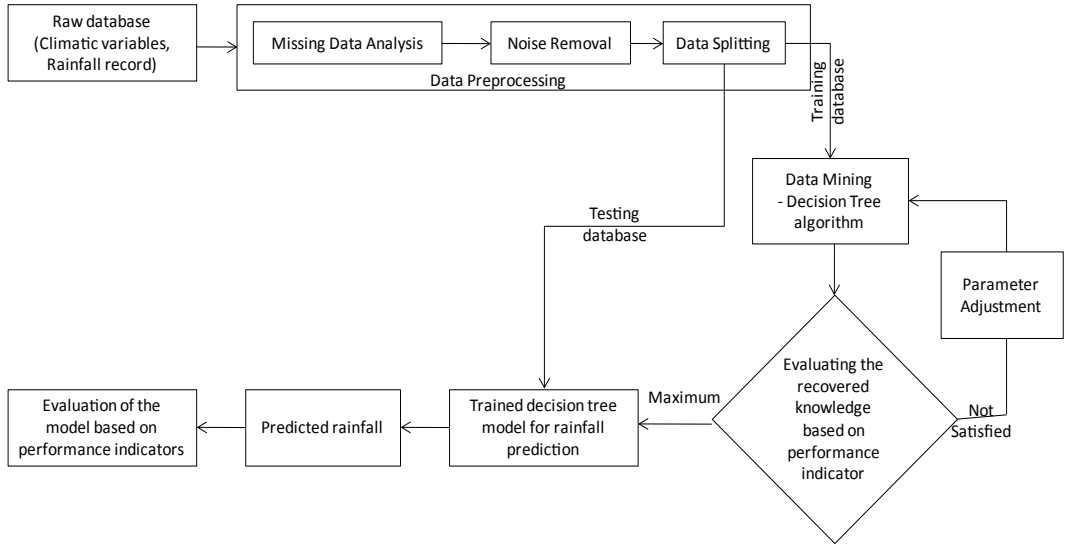

Figure 1: Decision tree based rainfall prediction model.

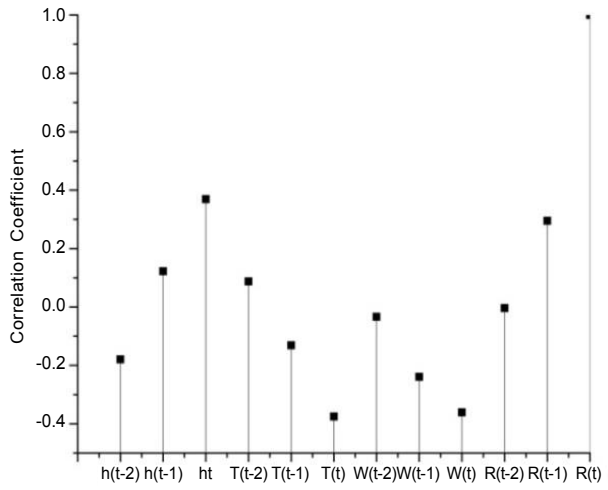

Figure 2: Correlation analysis for input variable selection.

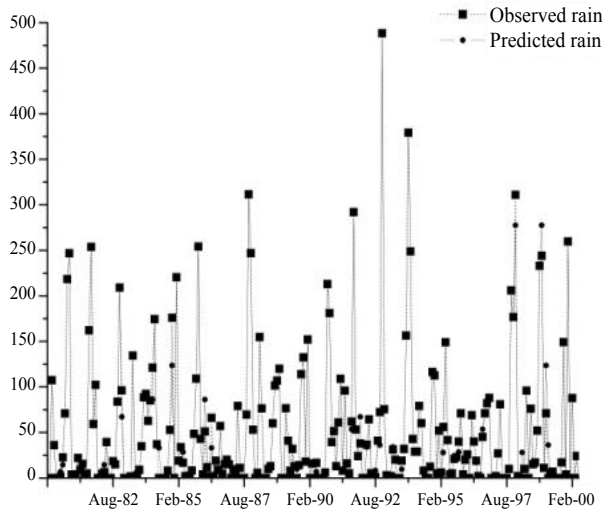

Figure 3: Comparison of observed and predicted rainfall during training phase.

propagation neural networks, consists of three distinct layers i) input layer, ii) hidden layer & iii) output layer [14-16]. A network consists one input, one or more hidden layers and one output layer. Information enters a network through the nodes of the input layer. The input layer nodes are unique in that their sole purpose is to distribute the input information to the next processing layer (i.e., the first hidden layer). This layer makes a computational analysis and the results can be seen in the output layer.

Application

The developed rainfall prediction framework has been applied to explore the climatic database of Tuticorin meteorological station, Tamil Nadu, India. The historical database had monthly record on humidity, temperature, wind speed and rainfall from year 1980 to 2002 [17]. The historical database has been preprocessed as highlighted in the methodology discussion. The input variables are selected as temperature at time‘t’, humidity at ‘t’, and rainfall at time ‘t-1’ based on correlation analysis (Figure 2).

Results and Discussion

The prepared database has been explored using DT data-mining algorithm for recovering the hidden relationship that exists between the input and output variables. Figure 3, shows the predicted rainfall by the developed RPD model during training phase. From Figure 3 and Table 1, it can be observed that the DT algorithm is able to recover the knowledge with R2 of 0.99 and Nash Sutcliffe efficiency (E) of 0.98, which highlights that the RPD is able to match the observed rainfall during the corresponding time periods compared to that of ANN based data-driven model. The recovered knowledge is used to predict the future rainfall events for 3 years. The RMSE and MBE are minimum in DT when compared to ANN model. The better results are observed in DT. ANN fails to map the observed rainfall with the predicted values.

Figure 4 summaries the efficiency of the RPD model in predicting the observed rainfall events during various time periods. From Figure 4, it

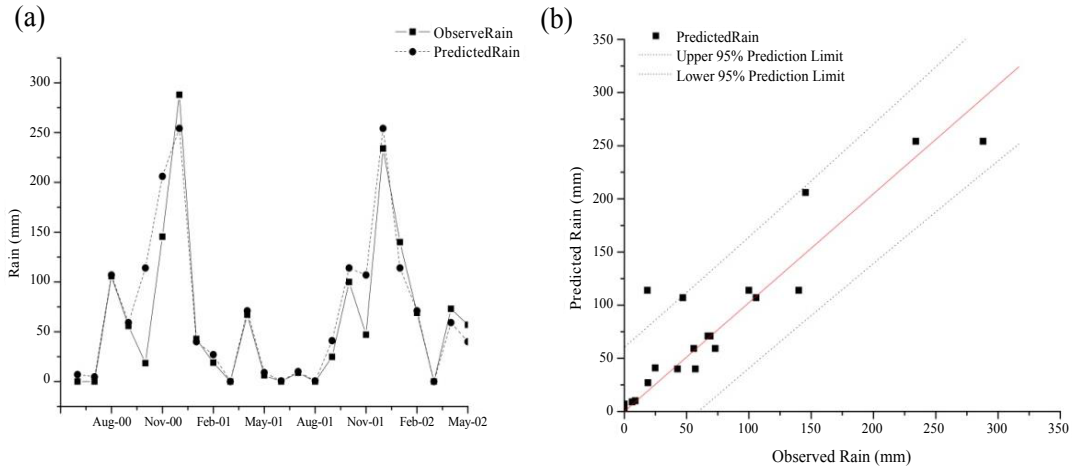

Figure 4: Comparison of observed and predicted rainfall during testing phase.

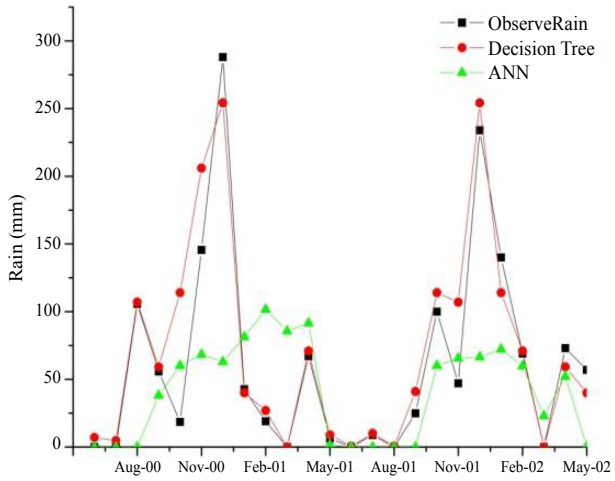

Figure 5: Comparison of and ANN and Decision tree predicted rainfall.

might be observed that the most of the predictions lies within the 95% prediction limit and only one outlier; this highlights reliability on the developed methodological framework.

Figure 5 compares the rainfall predictions of DT and ANN, it might be observed that DT outperforms in predicting the future rainfall events compared to ANN rainfall prediction model. From Table 1, it is inferred that ANN suffers to map the observed rainfall without most correlated input variables. However, DT data-mining algorithm is able to recover the most useful information within the available database and utilized the same to predict the future rainfalls with Nash Sutcliffe efficiency of 0.85.

Conclusions

In this research a Decision tree data-mining based rainfall prediction model has been developed to recover the knowledge from the historical database and utilize the same for future rainfall events. From the case study application, it may be observed that the developed RPD model is able to predict the future rain events comparatively better than ANN with available poorly correlated input variables. Further, this data-mining based rainfall prediction modeling may be improvised by increasing the length of historical database and also incorporating most correlated variables for modeling.

	Training		Testing	
	DT	ANN	DT	ANN
R^2	0.9932	0.5198	0.9328	0.3755
RMSE (mm)	8.8826	84.4973	28.6542	72.3478
MBE (mm)	2.7695	62.8109	16.4625	47.5025
E	0.9864	0.2119	0.8506	0.0486

Table 1: Performance of the rainfall prediction models during training and testing phase.

References

1. Cios KJ, Pedrycz W, Swiniarski RW, Kurgan IA (2010) Data-Mining: A knowledge discovery approach. Springer. New york.
2. Sudha V, Ambujam N, Venugopal K (2006) A data mining approch for deriving irrigation reservoir operating rules. Water Observation and Information System for decision Support, Macedonia.
3. Sattari MT, Apaydin H, Ozturk F, Baykal N (2012) Application of a data mining approach to derive operating rules for the Eleviyan irrigation reservoir. Lake and Reservoir Management. 28: 142-152.
4. Hernández S, Nesic S, Weckman G, Ghai V (2005) Use of artificial neural networks for predicting crude oil effect on carbobdioxide corrosion of carbon steels. Corrosion 2005. Paper No 05554.
5. Mellit A, Benghanen M, Kalogirou SA (2006) An adaptive wavelet network model for forecasting daily total solar radiation. Applied Energy. 83: 705-722.
6. Negm AM, Ibrahim AA, El-Saiad AA, Al-Brahim AM (2004) Flow resistance due to cylindrical piles. Egyptian Journal of Engineering Science and Technology. 7: 123-234.
7. Azimian AR (2005) Application of artificial neural networks in pipe flow calculations. 4th International Conference on Heat Transfer, Fluid Mechanics and Thermodynamics. Egypt. September 19-22.
8. Agrawal R, Imielinski T, Swami A (1993) Mining Association Rules between Sets of Items in Large Database. ACM SIGMOD Conference on Management of Data. Washington.
9. Chang LC, Chang FJ (2001) Intelligent control for modelling or real-time reservoir operation. Hydrological Processes. 15: 1621-1634.
10. Chaves P, TsukataniT, Kojiri T (2004) Operation of storage reservoir for water quality by using optimization and artificial intelligence techniques. Mathematics and Computers in Simulation. 67: 419-432.
11. Dawson CW, Wilby R (1998) An artificial neural network approach to rainfall-runoff modelling. Hydrological Sciences Journal. 43: 47-66.

12. French MN, Krajewski WF, Cuykendall RR (1992) Rainfall forecasting in Space and Time Using a Neural Network. J. Hydrol. 137: 1-31.

13. Keskin M, Terzi Ö (2006) Artificial Neural Network Models of Daily Pan Evaporation. Journal of Hydrologic Engineering. 11: 65-70.

14. Jain S, Das A, Srivastava D (1999) Application of ANN for Reservoir Inflow Prediction and Operation. Journal of Water Resources Planning Management. 125: 263-271.

15. Shirsath PB, Singh AK (2010) A Comparative Study of Daily Pan Evaporation Estimation Using ANN, Regression and Climate Based Models. Water Resources Management. 24: 1571-1581.

16. Witten IH, Frank E (2005) data-mining: Practical Machine learning tools and Techniques. Morgan Kaufmann Publishers, Elsever.

17. Tokar A, Johnson P (1999) Rainfall-Runoff Modeling Using Artificial Neural Networks. Journal of Hydrologic Engineering. 4: 232-239.

Citation: Ramsundram N, Sathya S, Karthikeyan S (2016) Comparison of Decision Tree Based Rainfall Prediction Model with Data Driven Model Considering Climatic Variables. Irrigat Drainage Sys Eng 5: 175. doi: 10.4172/2168-9768.1000175