International Journal of Research in Advent Technology, Vol.2, No.2, February 2014

E-ISSN: 2321-9637

A Data Mining Approach Using Artificial Neural Network to Predict Indian Monsoon Rainfall

Amruta A.Taksande, Dr.S.P.Khandait, Prof.Manish Katkar

Department of CSE

Email:ttejswini71@gmail.com

Abstract- Rainfall forecasting is very challenging task for everyone because it consists of multidimensional and nonlinear data. This paper describes five data mining algorithms namely neural network (NN), random forest, classification and regression tree (CRT), support vector machine (SVM) and *k*-nearest neighbour. Generally these algorithms are used for the prediction. But there are some limitations in these algorithms such as neural network as it is complex relationship between input and output variables .Therefore we use genetic algorithm for future prediction. This paper proposes neural network for predict the weather information such as pressure, temperature, humidity etc. and genetic algorithm for collecting and saving these data for the future prediction.

Index Terms- Data Mining Algorithms, Summer Monsoon Prediction, Artificial Neural Network, Genetic Algorithms, and Weather Forecasting.

1. INTRODUCTION

Weather forecasting is the application of science and technology to predict the state of the atmosphere for a given location. Human beings have attempted to predict the weather informally for millennia, and formally since the nineteenth century. Weather forecasts are made by collecting quantitative data about the current state of the atmosphere on a given place and using scientific understanding of atmospheric processes to project how the atmosphere will evolve on that place. Weather forecasting is one of the most imperative and demanding operational responsibilities carried out by meteorological services all over the world. The term “monsoon” seems to have been derived either from the Arabic mausin or from the Malayan monsin.[1].As first used it was applied to southern Asia and the adjacent waters, where it referred to the seasonal surface air streams which reverse their directions between winter and summer, southwest in summer and northeast in winter in this area. In 1686 Halley explained the Asiatic monsoon as resulting from thermal contrasts between the continent and oceans. Rainfall is one of several important factors affecting watershed water quality [1]. The downstream flux of nitrogen and phosphorus originating in the watershed basin depends on the amount of rainfall. Once an all-human endeavor based mainly upon changes in barometric pressure, current weather conditions, and sky condition, weather forecasting now relies on computer-based models that take many atmospheric factors into account. Present paper endeavors to develop an Artificial Neural Network (ANN) model to forecast average rainfall during summer-monsoon in India.

Human input is still required to pick the best possible forecast model to base the forecast upon, which involves pattern recognition skills, tele connections, knowledge of model performance, and knowledge of model biases. [3]The chaotic nature of the atmosphere, the massive computational power required to solve the equations that describe the atmosphere, error involved in measuring the initial conditions, and an incomplete understanding of atmospheric processes mean that forecasts become less accurate as the difference in current time and the time for which the forecast is being made (the *range* of the forecast) increases. There are a variety of end uses to weather forecasts. Weather warnings are important forecasts because they are used to protect life and property. Forecasts based on temperature and precipitation are important to agriculture, and therefore to traders within commodity markets. Temperature forecasts are used by utility companies to estimate demand over coming days. On an everyday basis, people use weather forecasts to determine what to wear on a given day. Since outdoor activities are severely curtailed by heavy rain, snow and the wind chill, forecasts can be used to plan activities around these events, and to plan ahead and survive them. Thus, rainfall forecasting can warn of happening flood or drought so that peoples can save their lives and properties. Rainfall forecasting is also important for engineering applications, mainly for the design of hydroelectric power projects, because this system requires prior information about average rainfall, maximum/minimum rainfall, maximum intensity, duration etc. for a year/each month. Thus, we believe that precise rainfall prediction is important for practitioners who are interested to make wise policies related to this event. In this paper we implement the Neural network, Genetic algorithm and Hidden Markov Model for the future prediction and optimization [4].

2. A BRIEF LITERATURE REVIEWS

In paper [1], five data-mining algorithms, neural network, random forest, classification and regression tree, support vector machine, and *k*-nearest neighbor were used to build the prediction models.

In paper [2], studied the sensitivity of satellite retrieval and sampling error on flood prediction uncertainty for a week-lasting rainfall event over a medium-sized watershed of a typical sensor footprint size (~ 100 km). Runoff uncertainty was found sensitive to both, systematic and random error components of PM retrievals.

Paper [3], uses a totally different approach, namely, a neural network based technique, is introduced to address the rainfall estimation problem using radar data. The neural network directly maps the radar observations to rainfall on the ground.

This paper [4], proposes a novel neural network (NN) training method that employs the hybrid exponential smoothing method and the Levenberg–Marquardt (LM) algorithm, which aims to improve the generalization capabilities of previously used methods for training NNs for short-term traffic flow forecasting. The proposed method was evaluated by forecasting short-term traffic flow conditions on the Mitchell freeway in Western Australia. With regard to the generalization capabilities for short-term traffic flow forecasting, the NN models developed using the proposed approach outperform those that are developed based on the alternative tested algorithms, which are particularly designed either for short-term traffic flow forecasting or for enhancing generalization capabilities of NNs.

3. DATA MINING ALGORITHMS

Five data-mining algorithms, neural network (NN), random forest, classification and regression tree (C&RT), support vector machine (SVM), and *k*-nearest neighbour (*k*-NN), were used to build the prediction models[1]. NN is a computational model which is inspired by the brain. NN consists of a group of interconnected neurons, making it an adaptive system that can change its structure based on external or internal information flowing through the network during the learning phase. NNs are usually used to model complex relationships between input and output variables. In this paper, an NN model known as the multilayer perception (MLP) is used [1][3].

Random forest combines decision tree predictors in a way that each tree depends on the values of a random vector sampled independently and with the same distribution for all trees in the forest. It integrates a bagging idea and a random selection of features in constructing a collection of decision trees. The algorithm is efficient for large data sets and is able to maintain good accuracy when a large proportion of the data is missing.

C&RT, popularized by Breiman, is a nonparametric technique producing logical if–then rules that are easy to interpret.

An SVM is a supervised learning method used for classification and regression analysis. SVM constructs one or a set of hyper planes in a high or infinite dimensional space. The data points in the feature (variable) space are mapped into the
Hyper plane by selected kernel functions. The key advantage of SVM is the use of kernel functions making SVM suitable for modelling in complex nonlinear domains.

k-NN is an instance-based learning method accounting for contributions of the neighbours. The nearest neighbours contribute more to the computed average than the distant ones. *k*- NN is simple and easy to implement. It offers good performance for some classes of applications.

4. ADVANCED METHODS THAT CAN BE USED FOR THE PREDICTION

A) Neural Network for Rainfall Prediction

Rainfall rate obtained on the ground can be potentially dependent on the 3-D structure of precipitation aloft [3][1]. In principle one can try to obtain a functional approximation between rainfall on the ground and the 3-D radar observation above the observation point. Therefore, the rainfall estimation problem can be viewed as a complex function approximation problem.

The universal approximation theorem for neural network states that a two-layer feed forward perceptron network with non constant, bounded, and monotone-increasing continuous activation function can perform arbitrary nonlinear input-output relationship mapping . A neural network learns the input–output relationship through the training process [1]. The learning process in a neural network is an interactive procedure in which its connection weights are adapted through the presentation of a set of input–output training example pairs. The gradient descent based back-propagation algorithm is the most popular learning algorithm for multilayer perceptrons (MLP). Therefore, a two-layer perceptron network can be used for the rainfall estimation problem. The above universal approximation theorem gives the theoretical justification for the approximation of an arbitrary continuous function by a two-layer (one hidden-layer) perceptron network. In practice, however, a three-layer (two hidden - layer) perceptron network works better than a two layer perceptron for the function approximation problem [3]. This is because the interaction between neurons in a single hidden layer network makes it difficult to obtain a globally good approximation, while a two-hidden layer network isolates and thus reduces the interaction effects by solving the problem in two steps, i.e., the first hidden-layer extracts the local features of the input data whereas the second hidden-layer extracts the global feature, to make the approximations in different regions of the input space individually adjusted . Due to above reasons, three-layer perceptron networks are chosen in this paper for the rainfall estimation problem. The structure of a three-layer perceptron is shown in Fig.1

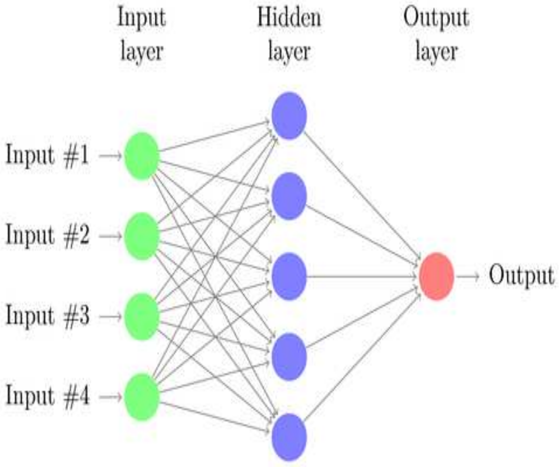

Fig 1. A Simple three layer ANN [5]

This paper develops ANN model to predict the average rainfall over India during summer- monsoon by exploring the data will be receiving from the yahoo weather API. The ANN approach has several advantages over conventional phenomenological or semi-empirical models, since they require known input data set without any assumptions. It exhibits rapid information processing and is able to develop a mapping of the input and output variables. Such a mapping can subsequently be used to predict desired outputs as a function of suitable inputs. A multilayer neural network can approximate any smooth, measurable function between input and output vectors by selecting a suitable set of connecting weights and transfer functions or activation function.

The model building process consists of four sequential steps:
(i) Selection of the input and output for the supervised Back propagation learning
(ii) Selection of the activation function
(iii) Training and testing of the model
(iv) Testing the goodness of fit of the model

B) Genetic Algorithm

This technique is proposed by Holland (1975). Later Koza (1992) used this approach (called his method genetic programming (GP)) to evolve programs to perform certain tasks. By definition, GA is a technique based on the 'Darwin's Principle of Natural Selection' and is used to solve optimisation problems. The basic idea is to select the best, discard the rest and differs from other systems in the following way. GA does not require any assumptions such as linearity, stationarity, homogeneity, chaotic or others. Thus, to handle the complex multi-dimensional behaviours of a system, this approach has been use deficiently in literature.

Special Features

Initialization:
Initially many individual solutions are (usually) randomly generated to form an initial population. The population size depends on the nature of the problem, but typically contains several hundreds or thousands of possible solutions.

- Selection:
During each successive generation, a proportion of the existing population is selected to breed a new generation. Individual solutions are selected through a

fitness-based process, where fitter solutions (as measured by a fitness function) are typically more likely to be selected.

- Genetic operators:
  The next step is to generate a second generation population of solutions from those selected through genetic operators: crossover (also called recombination), and/or mutation. For each new solution to be produced, a pair of "parent" solutions is selected for breeding from the pool selected previously. By producing a "child" solution using the above methods of crossover and mutation, a new solution is created which typically shares many of the characteristics of its "parents".
- Termination:
  This generational process is repeated until a termination condition has been reached. Common terminating conditions are:
  1. A solution is found that satisfies minimum criteria
  2. Fixed number of generations reached
  3. Manual inspection

5. CONCLUSION

In this paper we have presented a framework to develop neural network estimates of rainfall. Among the five data-mining algorithms tested in this paper, the MLP(multilayer perceptron) has performed best ANNs are being used increasingly for the prediction and forecasting of a number of water resources variables, including rainfall, flow, water level and various water quality parameters. In most papers, a good description of basic ANN theory, the case study considered and the results obtained is given.

This paper modelled the complex multi-dimensional behaviours of monthly monsoon rainfall for a number of stations using several soft computing techniques.

In this way we are collecting the weather information from yahoo weather API, the information such as pressure, temperature, humidity etc. These information will be stored as a database server for the future prediction. Our future research plan is to model the daily monsoon rainfall data, which have significance in the field of agriculture, transportation, sports, tourism activities and others by using Genetic algorithm (GA) and Hidden Markov Model (HMM).

REFERENCES

[1] Andrew Kusiak.(2013):“Modelling and Prediction of Rainfall Using Radar Reflectivity Data: A Data-Mining Approach” IEEE trans on geo science and remote sensing, VOL.51, NO.4,pp.2238-2239.

[2] Rongrui Xiao;Member; Chandrashekar V; Member, IEEE(2007): “Development of a Neural Network Based Algorithm for Rainfall Estimation from Radar Observations” IEEE trans on geo science and remote sensing,VOL.35,NO.1,pp.160.

[3] Kit Yan Chan;(2012): “Neural-Network-Based Models for Short-Term Traffic Flow Forecasting Using a Hybrid Exponential Smoothing and Leven berg–Marquardt Algorithm” IEEE trans on intelligent transportation system, VOL. 13, NO. 2, pp.644-646.

[4] Stephen Dunne;Bidisha Ghosh;(2013): “Weather Adaptive Traffic Prediction Using Neuro wavelet Models” , IEEE trans on intelligent transportation system, VOL. 14, NO. 1,pp.370.

[5] Dezhi Li;Wilson Wang; Fathy Ismail ;(2013):“Fuzzy Neural Network Technique for System State Forecasting” IEEE TRANSACTIONS ON CYBERNETICS, VOL. 43, NO. 5.

[6] Carlos Domenech;Tobias Wehr;(2011): “Use of Artificial Neural Networks to Retrieve TOA SW Radiative Fluxes for the Earth CARE Mission” IEEE trans on geo science and remote sensing,VOL.49,NO.6,pp.1841-1843.

[7] Fauvel M.; Benediktsson J.A.; Chanussot; Sveinsson J.R.;(2008): “Spectral and spatial classification of hyperspectral data using SVMs and morphological profiles,” IEEE Trans. Geosci. Remote Sens., vol. 46, no. 11,pp. 3804–3814.