JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue
JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)
An International Scholarly Open Access, Peer-reviewed, Refereed Journal

WEKA Models for Rainfall Data

S. Damodharan[1], S. Venkatramana Reddy[2*]and B. Sarojamma[3]

1&3: Department of Statistics, Sri Venkateswara University, Tirupati – 517 502, A.P., India.

2: Department of Physics, Sri Venkateswara University, Tirupati – 517 502, A.P., India.

* Author for Correspondence e-mail: drsvreddy123@gmail.com

ABSTRACT:

Rainfall plays a vital role in India for drinking and irrigation processes. In India there are four seasons, according to seasonal adjustment they are winter in the months of January and February, summer is from March to May, monsoon or rainy season from June to September and post monsoon from October to December. Generally average annual rainfall in India is 1200 millimeters or 120 centimeters. In this paper, we are fitted models by using Rep tree, Additive regression, Random Sub Space and Decision Table using WEKA software. Which model is the best estimates using different measures of accuracy like, Root absolute square error (RASE), Relative absolute error(RAE), root relative squared error (RSE) and symmetric mean absolute percentage error (SMAPE).

Keywords: Rainfall, Rep tree, Additive regression, Random Sub Space, Decision Table and SMAPE

1. Introduction:

In India the rainy season is generally from June to September, and the annual average rain recorded between 750 and 1500 millimeters around the region. Usually, rainwater is better than artificial irrigation methods because it does not have any added chemicals such as chlorine. The advantages are that it is good for plants and soil. It helps in reducing runoff pollution, contribute to erosion prevention efforts, and eco-friendly options to keep composts moister. On the other side of the coin are essential created flooding kills thousands around the world every year. Rain causes excessive load on the drainage system.

W.F. Krejewski et al [1] explains "Radar hydrology: rainfall estimation", the authors used Radar observations of rainfall and discussed their use. Methodological advances are needed in several areas of radar-rainfall estimation of particular importance, advance in rainfall estimates using polarimetric radar observations, estimates of the error structure of rainfall rate estimates and validation of radar rainfall algorithm. J.H.C.Gash[2] studies "An analytical model of rainfall interception by forecasts", and the two major factors which control the evaporation of intercepted rainfall area a) amount of time that the canopy spends saturated during rains and the evaporation rate applicable under these conditions and b) canopy saturates capacity and the number of times this store is emptied.

Jan G De Gooijer, et al [3] gives 25 years of time series forecasting and they explained exponential smoothing models, ARIMA models, Seasonality, State space and structural models, the Kalman filter, Non-linear, long memory ARCH/GARCH, current data forecasting and also explain different accuracy measures for forecasting prediction intervals and densities. M.Sidiq[4], in his article "Forecasting Rainfall with Time Series Model," aims to study the forecast rainfall with the time series model. For computations of ARIMA (1,0,1), ARIMA (0,1,1), ARIMA(1,1,1), AR(1) and MA(1) for months data from 2011 to 2014, i.e. 48 data points of banding in millimeters. They prefer ACF and PACF plots and calculations of AIC, MAE, and MAPE to choose a good model. D.N. Gujarati [5] gives the Essential of econometrics for different econometric models. A. Nogroho et al [6] explains "ARMA model for prediction of rainfall in Regency of semarany-central Java-Republic of Indonesia" and discussed different ARMA models for rainfall data.

2. Methodology:

Weather Forecasting is one of the greatest problems faced by Meteorology department. The data for weather can be synoptic according to the need, factors like Accuracy, Precision and others has very inevitable role in Forecasting. In Classification, the main objective is to anticipate the target class by examining the training dataset. A classification model is judged by putting it to experimental data with known output values and contrasting the predicted values with known values. Generally, the build data and experimental data come from the same former dataset. A small piece of the data is used to build model while the remaining records are used to test model. Now the Precision refers to the proportion of correct Predictive analysis made by the model. Then the Confusion Matrix displays the correct and incorrect number of predictions made by the model. It is an alternative way to check the Accuracy. However, we are considering the best performance according to Metrics such as Accuracy, Kappa statistic, RMSE, MAE and SMAPE

Weather Forecasting is done by powerful Super Computers which process hundreds of thousands of observations of current weather conditions. The data which we get from satellites is in Raw format which doesn't provide any kind of information. Therefore, to get knowledge, we need to process it using various mathematical models. The process of converting Raw data into information is known as Data Mining. Further, to predict the information various methods of Data Mining are used such as

- Decision Trees
- Rule based method
- Neural Networks
- Naïve Bayes
- Bayesian Belief Network
- Support Vector Machine

Among these methods, the most famous method is the Decision Tree method. The most widely used tool is WEKA (Waikato Environment for Knowledge Analysis). Weka is a collection of machine learning algorithms for Data Mining tasks. It contains tools for data preprocessing, Classification, Regression, Clustering, Association rule and Visualization.

Dataset and Preprocessing Classification methods:

Rep Tree:

Reptree Algorithm is a fast Decision tree Learner. It is also based on C4.5 Algorithm and can produce Classification or Regression trees. It builds a model using information and prunes it using reduced error pruning.

Additive Regression Tree:

In general additive regression algorithm is

$$Y_i = bo + b_1\ x_i + e_i$$

Where y_i is dependent variable

x_i is independent variable

e_i is error variable

b_o is intercept of regression line

b_1 is slope of regression lines

Random Sub Space:

The random subspace method is similar to bagging except that the features are randomly sampled, with replacement for each learner. Informally, this causes individual learners to not over-focus on features that appear highly predictive/descriptive in the training set, but fail to be as predictive for points outside that set. For this reason, random subspaces are an attractive choice for high-dimensional problems where the number of features is much larger than the number of training points, such as learning from fMRI data or gene expression data.

The random subspace method has been used for decision trees, when combined with "ordinary" bagging of decision trees, the resulting models are called random forests. It has also been applied to linear classifiers, support vector machines, nearest neighbors and other types of classifiers. This method is also applicable to one-class classifiers. Recently, the random subspace method has been used in a portfolio selection problem showing its superiority to the conventional resampled portfolio essentially based on Bagging.

Decision Table:

Decision Table is an accurate method for numeric prediction from decision trees and it is an ordered set of If-Then rules that have the potential to be more compact and therefore more understandable than the decision trees. Selection to explore decision tables because it is a simpler, less compute intensive algorithm than the decision-tree-based approach. The algorithm, decision table, is found in the Weka classifiers under Rules. The simplest way of representing the output from machine learning is to put it in the same form as the input. It summarizes the dataset with a "decision table‟ which contains the same number of attributes as the original dataset. The use of the classifier rules decision table is described as building and using a simple decision table majority classifier. The output will show a decision on a number of attributes for each instance. The number and specific types of attributes can vary to suit the needs of the task. Decision Table classifier algorithm is used to summarize the dataset by using a decision table containing the same number of attributes as that of the original dataset. A new data item is allocated a category by searching the line in the decision table that is equivalent to the values contained in the non-class of the data item. The entire problem of learning decision tables consists of selecting the right attributes to be included. Usually this is done by measuring the tables cross validation performance for different subsets of attributes and choosing the best performing subset. Fortunately, leave-one-out cross-validation is very cheap for this kind of classifier. Obtaining the cross-validation error from a decision table derived from the training data is just a matter of manipulating the class counts associated with each of the table's entries, because the table's structure doesn't change when instances are added or deleted. The attribute space is generally searched by best-first search because this strategy is less likely to get stuck in a local maximum than others, such as forward selection.

Accuracy Measurer: The various accuracy measures used in this paper are mean absolute error, root mean squared error, relative absolute error and root relative squared error.

Mean absolute Error: Difference between original value and estimated value error. By ignoring the sign of error values gives absolute error. Average of absolute error is mean absolute error.

Root mean squared error: difference between forecasted values and present values given error. Next step is square error values. Positive square root of squared error gives root mean square error.

Relative Absolute Error (RAE) : Relative absolute Error is very similar to relative squared error in relative absolute error, the error is just the absolute error instead of the total squared error. Therefore, relative absolute error takes total absolute error of the simple predictor. If relative absolute error is denoted with Ei formulae becomes.

$$E_i \ `= \frac{\sum_{j=1}^{n} |F_{ij} - Y_j|}{\sum_{j=1}^{n} |Y_j - \bar{\bar{y}}|}$$

Root Relative Squared Error(RRSE) : The positive square root of relative squared error takes the total squared error and normalizes it by dividing by total squared error of the simple estimates. The formula for RRSE is.

$$RRSE = \sqrt{\frac{\sum_{j=1}^{n}(P_{ij}-Y_j)^2}{\sum_{j=1}^{n}(Y_{j}-\overline{Y_j})^2}}$$

Symmetric mean absolute percentage error (**SMAPE**): **Symmetric mean absolute percentage error (SMAPE** or **sMAPE)** is an accuracy measure based on percentage (or relative) errors. It is usually defined as follows:

$$SMAPE = \frac{1}{n}\sum_{i=1}^{n}\frac{|F_t-Y_t|}{(Y_t+F_t)/_2}$$

where Y_t is the actual value and F_t is the forecast value.

The absolute difference between Y_t and F_t is divided by half the sum of absolute values of the actual value Y_t and the forecast value F_t. The value of this calculation is summed for every fitted point *t* and divided again by the number of fitted points *n*.

3. Empirical investigations:

We are used annual rainfall data of india from 1992 to 2016[7]. By using WEKA software, we are produced Reptree, Additive Regression, Random Sub space and decision Table for fitting and prediction of rain fall data. The below table explains the predicted values and listed their error as follows.

		reptree	reptree	Additive regression	Additive regression	random sub space	random sub space	decision table	decision table
year	annual rainfall	predicted,	error	predicted,	error	predicted,	error	predicted,	error
1992	5682	5876.995	194.995	6165.901	483.901	5876.995	194.995	5663.364	-18.636
1993	6305	3753.667	-2551.333	5959.294	-345.706	3753.667	-2551.333	5663.364	-641.636
1994	5543	3753.667	-1789.333	4329.994	-1213.006	3753.667	-1789.333	5663.364	120.364
1995	6508	5674.945	-833.055	5968.969	-539.031	5674.945	-833.055	5692.227	-815.773
1996	4992	5674.945	682.945	6085.904	1093.904	5674.945	682.945	5692.227	700.227
1997	5395	5674.945	279.945	5026.773	-368.227	5674.945	279.945	5692.227	297.227
1998	4862	4861.438	-0.563	5552.851	690.851	4861.438	-0.563	5848.727	986.727

1999	4163	4861.438	698.438	5963.938	1800.938	4861.438	698.438	5848.727	1685.727
2000	4427	6544	2117	6191.168	1764.168	6544	2117	5848.727	1421.727
2001	6752	5581.037	-1170.963	5537.374	-1214.626	5581.037	-1170.963	5564.591	-1187.409
2002	7369	5581.037	-1787.963	5441.342	-1927.658	5581.037	-1787.963	5564.591	-1804.409
2003	5582	5581.037	-0.963	6125.432	543.432	5581.037	-0.963	5564.591	-17.409
2004	5633	5494.158	-138.842	5515.16	-117.84	5494.158	-138.842	5549.682	-83.318
2005	6619	5494.158	-1124.842	5965.097	-653.903	5494.158	-1124.842	5549.682	-1069.318
2006	7779	5494.158	-2284.842	5237.5	-2541.5	5494.158	-2284.842	5549.682	-2229.318
2007	6044	5741.69	-302.31	4820.775	-1223.225	5741.69	-302.31	5744.043	-299.957
2008	3967	5741.69	1774.69	4820.775	853.775	5741.69	1774.69	5744.043	1777.043
2009	7724	5541.533	-2182.467	5787.333	-1936.667	5541.533	-2182.467	5607.826	-2116.174
2010	5420	5541.533	121.533	5632.087	212.087	5541.533	121.533	5607.826	187.826
2011	5404	5720.413	316.413	5609.983	205.983	5720.413	316.413	5716.87	312.87
2012	5232	5720.413	488.413	7057.011	1825.011	5720.413	488.413	5716.87	484.87
2013	6574	5683.861	-890.139	6492.912	-81.088	5683.861	-890.139	5680.696	-893.304
2014	4894	5683.861	789.861	6962.436	2068.436	5683.861	789.861	5680.696	786.696
2015	6123	5725.047	-397.953	5994.7	-128.3	5725.047	-397.953	5776.957	-346.043
2016	3131	5725.047	2594.047	5275.002	2144.002	5725.047	2594.047	5776.957	2645.957

In the above table first column explains years from 1993 to 2016, second column describes about annual rainfall values in millimeters, third column explains about predicted values of rainfall using reptree, fourth column says about predicted observations of annual rainfall using WEKA software, fifth column talks about predicted values of random sub space and sixth column tells about decision table predicted values using WEKA.

The accuracy measures like Mean Absolute Error, Root Mean Squared Error, Relative Error and Root Relative Squared Error and correlation coefficient is as follows:

For Rep tree

Correlation coefficient	-0.0828
Mean absolute error	1020.554
Root mean squared error	1321.991
Relative absolute error	111.2686
Root relative squared error	111.8474
SMAPE	0.18664
Total Number of Instances	25

The accuracy measures like Mean Absolute Error, Root Mean Squared Error, Relative Error and Root Relative Squared Error and correlation coefficient is as follows for Additive Regression.

Column1	Column2
Correlation coefficient	-0.0021
Mean absolute error	1039.0906
Root mean squared error	1280.0022
Relative absolute error	113.2896
Root relative squared error	108.2949
SMAPE	0.1831
Total Number of Instances	25

The accuracy measures like Mean Absolute Error, Root Mean Squared Error, Relative Error and Root Relative Squared Error and correlation coefficient is as follows for Random Sub Space

Correlation coefficient	-0.0828
Mean absolute error	1020.554
Root mean squared error	1321.9911
Relative absolute error	111.2686
Root relative squared error	111.8474
SMAPE	0.1866
Total Number of Instances	25

The accuracy measures like Mean Absolute Error, Root Mean Squared Error, Relative Error and Root Relative Squared Error and correlation coefficient is as follows for Decision Table.

Correlation coefficient	-0.6748
Mean absolute error	917.1987
Root mean squared error	1181.9594
Relative absolute error	100
Root relative squared error	100
SMAPE	0.1809
Total Number of Instances	25

4. Summary and conclusions:

Rain fall plays vital role in India for irrigation and drinking purpose. In this paper, we are fitted models by using Rep tree, Additive regression, Random Sub Space and Decision Table using WEKA software. Which model is the best estimates using different measures of accuracy like, Root absolute square error (RASE), Relative absolute error (RAE), root relative squared error (RRSE) and symmetric mean absolute percentage error (SMAPE).

model	Rep tree	Additive regression,	Random Sub Space	Decision Table
Correlation coefficient	-0.0828	-0.0021	-0.0828	-0.6748
Mean absolute error	1020.554	1039.0906	1020.554	917.1987
Root mean squared error	1321.991	1280.0022	1321.9911	1181.9594
Relative absolute error	111.2686	113.2896	111.2686	100
Root relative squared error	111.8474	108.2949	111.8474	100
SMAPE	0.18664	0.1831	0.1866	0.1809
Total Number of Instances	25	25	25	25

The best model fitted for rainfall data using WEKA software is Decision Table because of least SMAPE and RMSE values

References:

1. W.F.Krejewski, J.A.Smith, "Radar hydrology: rainfall estimation", Advances in water Resourses 25 (2002)1387-1394.
2. J.H.C.Gash, "An analytical model of rainfall interception by forests", Quarterly Journal of the Royal metrological society, 105(1979)43-45.
3. Jan G De Gooijer, "25 years' time series forecasting", International Journal of forecasting, 22(3) (2006)443-473.
4. M.Sidiq, "Forecasting rainfall with time series model", 10p conference series: Mathematical Science and Engineering, 407(2018)012154
doI:10.1.1088/1757-899X/407/1/012154.
5. D.N. Gujarati, "Essential of Econometrics", New York, MC Graw –Hill co. 2009.
6. A. Nugroho and B.H. Simanjuntak, "ARMA (Auto Regressive Moving Average) model for prediction of rainfall in the regency of Semarang-central Java-Republic of Indonesia", International Journal of Computer Science (IJCSI), 11(3) (2014)27.
7. Data website: URL:http://www.imd.gov.in/Welcome%20To%20IMD/Welcome.php