Predictive Data mining And Modified Kmeans For Rainfall Prediction

Dr.D.Devakumari[1], T.Angamuthu[2]
[1]Assistant Professor, Dept of Computer Science
[2]Dept of Computer Science
[1, 2] Government Arts College, Coimbatore-641018

Abstract- *In the present article, an attempt is made to derive optimal data-driven machine learning methods for forecasting an average daily and monthly rainfall of the Kaggle Dataset. This comparative study is conducted concentrating on three aspects: modelling inputs, modelling methods and pre-processing techniques. A comparison between linear correlation analysis and average mutual information is made to find an optimal input technique. For the modelling of the rainfall, a novel hybrid multi-model method is proposed M-kmeans and compared with its constituent models. The models include the Kmeans, multivariate adaptive regression splines, the k-nearest neighbour, and radial basis support vector regression. the MODIFIED K-MEANS showed higher classification accuracy (91.40%) over Kmeans (90.90) models.*

Keywords- rainfall forecasting, machine learning, multi-model method, pre-processing, model ranking.

I. INTRODUCTION

Accurate forecasting of rainfall has been one of the most important issues in hydrological research because early warnings of severe weather can help prevent casualties and damages caused by natural disasters, if timely and accurately forecasted. To construct a predictive system for accurate rainfall, forecasting is one of the greatest challenges to researchers from diverse fields such as weather data mining (Yang et al., 2007), environmental machine learning (Hong, 2008), operational hydrology (Li and Lai, 2004), and statistical forecasting (Pucheta et al., 2009). A common question in these problems is how one can analyse the past and use future prediction. The parameters that are required to predict rainfall are enormously complex and subtle even for a short term period. Physical processes in rainfall are generally composed of a number of sub-processes. A accurate modelling of rainfall by a single global model is sometimes not possible (Solomatine and Ostfeld, 2008). To overcome this difficulty, the concept of modular modelling and combining different models has attracted more attention recently in rainfall forecasting. In modular models, several sub-processes are first identified, and then separate models (also called local or expert models) are established for each of them (Solomatine and Ostfeld, 2008). So far, various modular models have been proposed, depending on soft or hard splitting of training data. Soft splitting means that the dataset can be overlapped, and the overall forecasting output is the weighted average of each local model (Shrestha and Solomatine, 2006; Wu et al., 2008). In the hard splitting, there is no overlap of data and the final forecasting output is derived explicitly from only one of the local models (Wu et al., 2008). The approach of combining several models is also known as ensemble modelling. The basic idea behind the ensemble model is to build several different models for the same process and to integrate them together. The Empirical approach is based on analysis of past historical data of weather and its relationship to a variety of atmospheric variables over different parts of Chhattisgarh. The most widely use empirical approaches used for climate prediction are Regression, artificial neural network, fuzzy logic and group method of data handling.

The dynamical approach, predictions are generated by physical models based on system of equations that predict the future Rainfall. The forecasting of weather by computer using equations are known as numerical weather prediction. To predict the weather by numeric means, meteorologist has develop atmospheric models that approximate the change in temperature, pressure etc using mathematical equations. Regression is a statistical measure that attempts to determine the strength of the relationship between one dependent variable usually denoted by Y and a series of other changing variables known as independent variables. Regression model which contain more than two predictor variables are called Multiple Regression Model.

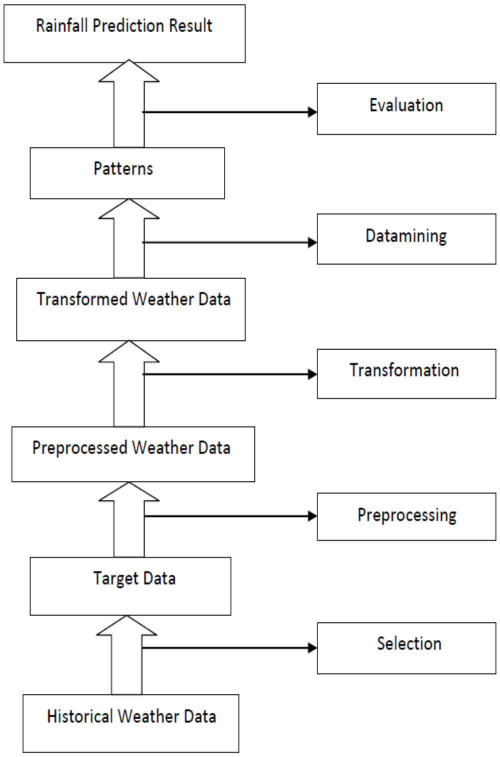


Figure1 forecasting for years of graph

The rest of the paper is organised as follows. In Section 2, we discuss briefly the study area and the rainfall series used in this paper. In Section 3, we describe the hybrid forecast model including the input selection technique and the variable selection method, and how the weights are extracted. This is followed by discussions about the experimental setup (Section 4) and results (Section 5). Conclusive discussions of the paper appear in Section 6.

II. LITRATURE REVIEW

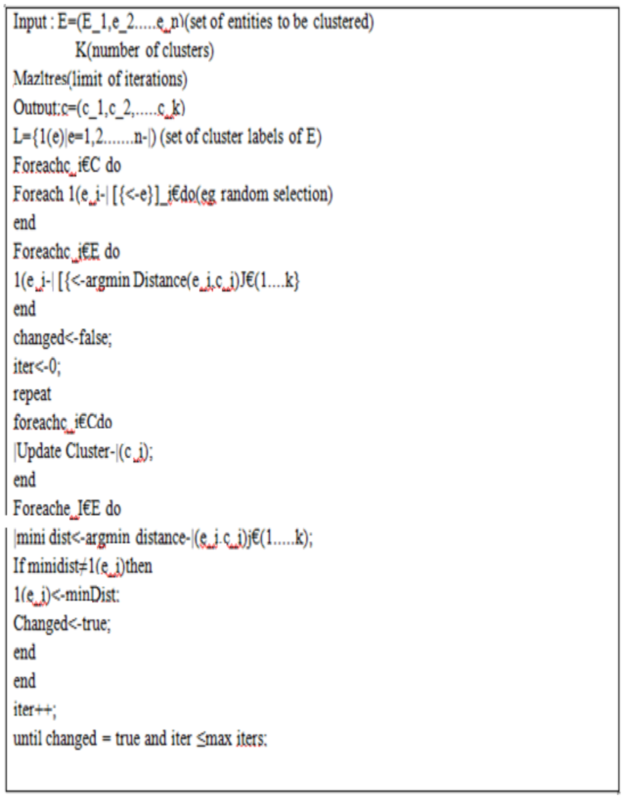
(Xiong et al., 2001; Abrahart and See, 2002; Kim et al., 2006; Baruque et al., 2011; Siwek et al., 2009; Zaman and Hirose, 2011). For example, Xiong et al. (2001) used a Takagi–Sugeno–Kang fuzzy technique to couple several conceptual rainfall-runoffmodels. Coulibaly et al. (2005) employed an improved weighted-average method to coalesce forecasted daily reservoir inflows from the k-Nearest Neighbor (k-NN), the conceptual model, and the Artificial Neural Network (ANN). Kim et al. (2006) investigated five ensemble methods for improving stream flow prediction.

The idea of ensemble learning is popular in other time series applications as well. Wichard and co-workers applied an ensemble of multi-models to construct hybrid models for NN5 time series competition (Wichard and Ogorzalek, 2007; Wichard, 2011). Deng et al. (2005) applied a parallel ensemble of support vector regression in two simulated time series datasets, the Sunspot and Mickey Glass datasets. A novel neural network ensemble approach called the generalized regression neural network ensemble for time series forecasting (GEFTSGRNN) which is a concatenation of existing machine learning algorithms has been applied in benchmark time series forecasting datasets by Gheyas and Smith (2011).

Everingham et al. (2009) constructed an ensemblemethod comprising statistical data mining models, to forecast crop productions in north eastern Australia. In this article, we make a comparison of several machine learning methods of forecasting an average daily and monthly rainfall of the Fukuoka city in Japan. All the methods are coupled with two data-preprocessing techniques. Prior to applying the methods, two input selection techniques are used. For the modelling of the rainfall, a novel hybrid multi-model method is proposed. The constituent models of the hybrid method are the ANN, Multivariate Adaptive Regression Splines (MARS), the k-nearest neighbour, and radial basis Support Vector Regression (SVR). The hybrid method generates sub-models first from each of the above methods with different parameter settings. Second, all the sub-models are ranked with a variable selection technique called least angle regression (LARS). Third, the higher ranked models are selected based on their Leave-One-Out Cross-Validation (LOOCV) error. The forecasting using the out of samples is done by a weighted combination (Timmermann, 2006) of the finally selected models. For evaluation of this hybrid method, we have constructed all these methods with their respective optimal parameters and applied to out of sample forecasting.

III. SYSTEM DESIGN

The proposed predictive model is used for the prediction of rain-fall. The predictive model is build using the available rainfall da-taset, mathematical equations and algorithms of data mining, ma-chine learning and so on. Very first step is, the dataset is prepro-cessed for removing unwanted data, noise, and finding the missing values. Once after preprocessing the data, the dataset is divided into two partitions like in training data and testing data i.e 80% of dataset is used for training purpose and 20% of data of dataset is used for the testing the predictive build model. Once after success-ful validation of the build model i.e the model working efficiently with correct output then the model is deployed for the future application.



```
Input : E=(E_1,e_2.....e_n)(set of entities to be clustered)
        K(number of clusters)
MazItres(limit of iterations)
Outout:c=(c_1,c_2,.....c_k)
L={1(e)|e=1,2.......n-|) (set of cluster labels of E)
Foreachc_i€C do
Foreach 1(e_i-| [{<-e}]_i€do(eg random selection)
end
Foreachc_i€E do
1(e_i-| [{<-argmin Distance(e_i,c_i)J€(1....k}
end
changed<-false;
iter<-0;
repeat
foreachc_i€Cdo
|Update Cluster-|(c_i);
end
Foreache_I€E do
|mini dist<-argmin distance-|(e_i.c_i)j€(1.....k);
If minidist≠1(e_i)then
1(e_i)<-minDist:
Changed<-true;
end
end
iter++;
until changed = true and iter ≤max iters:
```

Figure 2 psudocode for modified k-means algorithm

Algorithm:

The rainfall forecasting techniques or methods are presented as different algorithms.The working of each method and the signifi-cant operations are taken out.

Algorithm: Rainfall Prediction
Inputs: Rainfall dataset.
Output: Predicted Rainfall
Start
Step 1: import and accept rainfall dataset.
Step 2: Compute Avg. rainfall.
Step 3: Repeat Steps (3.1 to 3.4) for requested periods
Step3.1: Evaluate rainfall using linear fashion
Step3.2: If (Evaluated rainfall ≥ (Avg. rainfall + 10% Avg. Rain-fall)) then
Status ("Above Normal")
Step3.3: If (Evaluated Rainfall ≤ (Avg. rainfall - 10% Avg. Rain-fall)) then
Status ("Below Normal")
Step3.4: If (Evaluated rainfall lies between above and below nor-mal limit) then
Status ("Normal")
Step 4: Finally Display the Rainfall status.

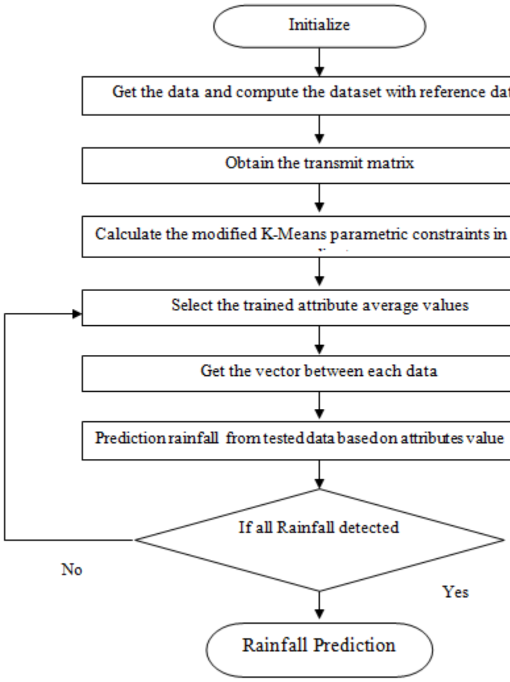

Initialize
Get the data and compute the dataset with reference data.
Obtain the transmit matrix
Calculate the modified K-Means parametric constraints in each
Select the trained attribute average values
Get the vector between each data
Prediction rainfall from tested data based on attributes value
If all Rainfall detected
No
Yes
Rainfall Prediction

Figure 3 Proposed modified K-Means model flow chart

IV. RESULT AND DISCUSSION

In this section, descripting the results in the form of graph using the R-Studio. By taking into the consideration of dataset of the annual rainfall in each subdivision of India from 1901 to 2015. The annual rainfall in each subdivision of India from 1901 to 2015 is shown in following figure. X-> Subdivision and Y-> Annual Rainfall in mm values. To analyze the performance of the proposed model, accuracy, precision and recall were used for evaluating classification results and mean squared error (MSE) and R2 score were used for evaluating regression results. The Modified K-Means Algorithm was chosen to solve this problem. Table 4.1 shows the Confusion matrix which is typically used to evaluate performance of data mining algorithms.

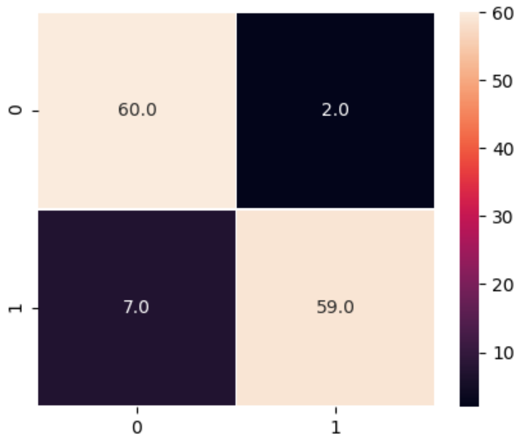
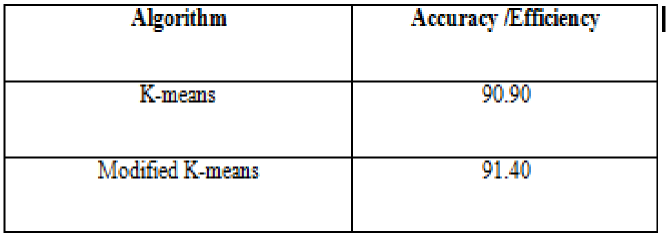


Figure 4 normalized MK-MEANS confusion matrix predicted values

Table1 Proposed Algorithm models compression with existing result



Algorithm	Accuracy /Efficiency
K-means	90.90
Modified K-means	91.40

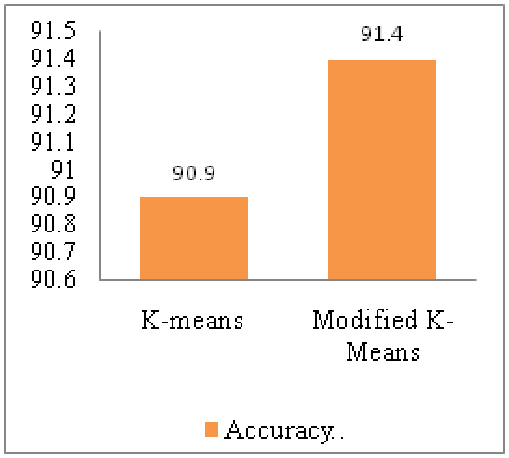


Figure 5 proposed Algorithm Accuracy compression results

V. CONCLUSION

The forecasting of rainfall is a very important factor in terms of water resource management, human life and their environment and prior to the agriculture for proper crop management. As rain-fall is a nonlinear in nature, it values are not constant, so statistical model yield poor inaccuracy in result. In the process of survey paper investigation of different prevalent Machine Learning, Data Mining and Satellite forecasting techniques and algorithm are presented to predict the rainfall. These techniques would help in predicting the accurate rainfall. However some limitations is clearly noticed in all the methods of rainfall prediction discussed in this survey paper The extensive references in support of the different developments of methods provided in this research should be of great help to solve their problem they will be facing in their proposed prediction model.

REFERENCES

[1] Abrahart, R.J. and See, L. (2002). Multi-model data fusion for river flow forecasting: An evaluation of six alternative methods based on two contrasting catchments, *Hydrology and Earth System Sciences* 6(4): 655–670.

[2] Baruque, B., Porras, S. and Corchado, E. (2011). Hybrid classification ensemble using topology-preserving clustering, *New Generation Computing* 29(3): 329–344.

[3] Chalimourda, A., Sch¨olkopf, B. and Smola, A.J. (2004). Experimentally optimal *ν* in support vector regression for different noise models and parameter settings, *Neural Networks: The Official Journal of the International Neural Network Society* 17(1): 127–41.

[4] Cherkassky, V. and Ma, Y. (2004). Practical selection of SVM parameters and noise estimation for SVM regression, *Neural Networks: The Official Journal of the International Neural Network Society* 17(1): 113–26.

[5] Coulibaly, P., Hach´e, M., Fortin, V. and Bob´ee, B. (2005). Improving daily reservoir inflow forecasts withmodel combination, Journal of Hydrologic Engineering 10(2): 91.

[6] Dawson, C.W. and Wilby, R.L. (2001). Hydrological modelling using artificial neural networks, Progress in Physical Geography 25(1): 80–108.

[7] De Vos, N.J. and Rientjes, T.H.M. (2005). Constraints of artificial neural networks for rainfall-runoff modelling: Trade-offs in hydrological state representation and model evaluation, Hydrology and Earth System Sciences 9(1–2): 111–126.

[8] Deng, Y.-F., Jin, X. and Zhong, Y.-X. (2005). Ensemble SVR for prediction of time series, Proceedings of the International Conference on Machine Learning and Cybernetics, Guangzhou, China, Vol. 2, pp. 734–748.

[9] Diebold, F.X. and Mariano, R.S. (1995). Comparing predictive accuracy, Journal of Business & Economic Statistics 13(3): 253–263.

[10] Efron, B., Hastie, T., Johnstone, I. and Tibshirani, R. (2004). Least angle regression, Annals of Statistics 32(2): 407–499. Everingham, Y.L., Smyth, C.W. and Inman-Bamber, N.G. (2009). Ensemble data mining approaches to forecast regional sugarcane crop production, Agricultural and Forest Meteorology 149(3–4): 689–696.

[11] Fraley, C. and Hesterberg, T. (2009). Least angle regression and LASSO for large datasets, Statistical Analysis and Data Mining 1(4): 251–259.

[12] Indrabayu1, Nadjamuddin Harun2, M. Saleh Pallu3, Andani Achmad4, " Numerical Statistic Approach for Expert System in Rainfall Prediction Based On Data Series" in International Journal of Computational Engineering Research,Vol, 0,Issue, 4,APRIL 2013.

[13] P. G. Popale and S.D. Gorantiwar Stochastic Generation and Forecasting Of Weekly Rainfall for Rahuri Region in International Journal of Innovative Research in Science, Engineering and Technology(IJIRSET) ISSN (Online) : 2319 – 8753 ISSN (Print) : 2347 – 6710 Volume 3, Special Issue 4, April 2014

[14] Saleh Zakaria, Nadhir Al-Ansari, Sven Knutsson, Thafer Al-Badrany "ARIMA Models for weekly rainfall in the semi-arid Sinjar District at Iraq" in Journal of Earth Sciences and Geotechnical Engineering, vol.2, no. 3, 2012, 25-55,ISSN: 1792-9040(print), 1792-9660 (online) Scienpress Ltd, 2012.

[15] K.Somvanshi, O.P.Pandey, P.K.Agrawal, N.V.Kalanker , M.Ravi Prakash and Ramesh Chand. "Modelling and prediction of rainfall using artificial neural network and ARIMA techniques" in J.Ind. Geophys.Union , Vol 10, No. 2, pp.141-151, April 2006.

[16] S.Meenakshi Sundaram, M.Lakshmi " Rainfall Prediction using Seasonal Auto Regressive Integrated Moving Average model" in- INDIAN JOURNAL OF RESEARCH ISSN - 2250-1991 Volume : 3 | Issue : 4 | April 2014.