Procedia Computer Science

Volume 51, 2015, Pages 2938–2942

ICCS 2015 International Conference On Computational Science

Co-Clustering Based Approach for Indian Monsoon Prediction

Moumita Saha and Pabitra Mitra

Department of Computer Science and Engineering, Indian Institute of Technology Kharagpur, India
moumita.saha2012@gmail.com, pabitra@gmail.com

Abstract

Prediction of Indian monsoon is a challenging task due to complex dynamics and variability over the years. Skills of statistical predictors that perform well in a set of years are not as good for others. In this paper, we attempt to identify a set of predictors that have high skills for a cluster of years. A co-clustering algorithm, which extracts groups of years, paired with good predictor sets for those years, is used for this purpose. Weighted ensemble of these predictors are used in final prediction. Results on past 65 years data show that the approach is competitive with state of art techniques.

Keywords: Indian monsoon prediction, co-clustering, ensemble methods

1 Introduction

Monsoon is influenced by a number of global climatic parameters. Variability of Indian monsoon is high and the correlation between Indian monsoon and predictor global climatic parameters changes over time. It is challenging to depict the variation as well as the predictor climatic parameters influencing Indian monsoon. Selection of separate predictor set for different clusters of monsoon years is essential, which is accomplished by co-clustering technique [1]. The method results in clustering the years of study into groups with distinct predictor climatic parameters for each group, which are important for prediction of Indian monsoon.

In our work, we first select global climatic parameters which are strongly correlated with Indian monsoon. It is followed by co-clustering of years into groups with different feature predictor climatic parameters. Prediction model is designed for each cluster for forecasting annual Indian rainfall. Finally, weighted ensemble of forecast given by respective models for each cluster is performed to acquire final forecast rainfall for the country. The proposed approach shows its imprint and ascertain its superiority in prediction of Indian monsoon.

Selection and peer-review under responsibility of the Scientific Programme Committee of ICCS 2015

© The Authors. Published by Elsevier B.V.

doi:10.1016/j.procs.2015.05.485

2 Co-Clustering Based Approach

Global Climatic Predictors of Indian Monsoon: A set of global predictors are selected from the studies made by Indian Meteorological Department (*IMD*) [3, 4]. Selected predictors include Equatorial Pacific Ocean sea level pressure (*SLP*)(*A1*), North Atlantic Ocean *SLP* (*A2*), East Asia surface pressure (*SP*) (*A3*), North West Europe *SP* (*A4*), North Atlantic Ocean sea surface temperature (*SST*) (*A5*), Equatorial South Eastern Indian ocean *SST* (*A6*), Indonesia *SST* (*A7*), North Central Pacific zonal wind (*A8*), Niño 3.4 *SST* (*A9*), pressure gradient between Madagascar and Tibetan low (*A10*), East Asia *SST* (*A11*), Europe land surface air temperature (*A12*), Madagascar *SLP* (*A13*), North Central Pacific Ocean *SLP* (*A14*), and Tibetan low *SP* (*A15*).

Data Sources: Sea surface temperature (*SST*) is obtained from *NOAA_OI_SST_V2* (*www.esrl.noaa.gov/ psd/*). Surface pressure (*SP*), air temperature, and zonal wind velocity are collected from *NCEP* Reanalysis Derived data (*www.esrl.noaa.gov/ psd/*). Niño 3.4 is acquired from National Center for Atmospheric Research (*www.climatedataguide.ucar.edu/climate-data/overview-climate-indices*). Sea level pressure (*SLP*) is obtained from 20^{th} Century Reanalysis data (*www.esrl.noaa.gov/psd/*). Annual Indian rainfall is collected from *Indian Institute of Tropical Meteorology, Pune* (*www.imdpune.gov.in/ research/ncc/longrange/ data/data.html*). All the monthly data are collected for time-period *1948-2012*.

Co-Clustering of Years and Predictors: Spectral co-clustering method [1] is used for dual clustering of years and predictors into homogeneous groups. It treats input data matrix as a bipartite graph and approximates the normalized cut by generalized eigen value decomposition of laplacian of the graph. Set of co-clusters are obtained, where each row and each column belongs to exactly one co-cluster.

3 Cluster Based Ensemble of Predictions

Separate models are built for every clusters of years with suitable predictor sets, which are obtained by co-clustering method. Models used are as following– (i) fitted ensemble of regression tree with bagging algorithm (*RegTreeB*), (ii) ensemble of bagged decision tree (*DecTreeB*), and (iii) multiple regression (*MultipleReg*). Final prediction is provided by weighted ensemble of predictions given by models designed for all clusters ($Pred_{final} = \sum_{i=1}^{c} W_i.Pi$, where $Pred_{final}$ is the final forecast, W_i and P_i are the weight assigned and prediction made by model built for i^{th} cluster, and c is the total number of clusters). Ensemble techniques used are as following.

- Simple arithmetic mean (*Ens1*): Equal weights are assigned to all models ($W_i = \frac{1}{c}$).
- Weighed ensemble in linear order (*Ens2*): Weight is taken as a function of distance of test year from the cluster center in linear order ($W_i = \frac{1}{d_i}$, where d_i is the euclidean distance of test year from cluster center of cluster i).
- Weighted ensemble in quadratic order (*Ens3*): Weight is assigned in quadratic order ($W_i = exp\left[\frac{1}{d_i}\right]$).
- Pearson correlation of past prediction (*Ens4*): Pearson correlation between predicted values by cluster model and actual value for the last 10 years is taken as weight for the prediction of the concurrent test year.

The schematic diagram of the proposed co-clustering-based ensemble method for prediction of Indian monsoon is shown in Figure 1.

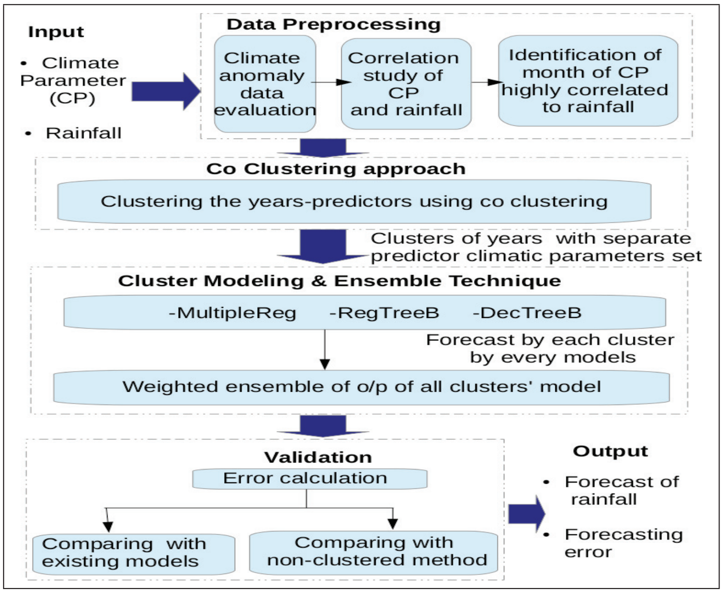
Input
• Climate Parameter (CP)
• Rainfall
Data Preprocessing
Climate anomaly data evaluation
Correlation study of CP and rainfall
Identification of month of CP highly correlated to rainfall
Co Clustering approach
Clustering the years-predictors using co clustering
Clusters of years with separate predictor climatic parameters set
Cluster Modeling & Ensemble Technique
-MultipleReg -RegTreeB -DecTreeB
Forecast by each cluster by every models
Weighted ensemble of o/p of all clusters' model
Validation
Error calculation
Comparing with existing models
Comparing with non-clustered method
Output
• Forecast of rainfall
• Forecasting error

Figure 1: Proposed co-clustering based ensemble approach for prediction of Indian monsoon

4 Experimental Results and Analysis

Dual Clustering of Predictors and Years: Indian summer monsoon rainfall occurring in months of June, July, August, and September, is considered for period *1948-2012* with *15* global predictor attributes (discussed in Section 2) for clustering. Co-clustering algorithm performs dual clustering of predictor set and years of study simultaneously. Number of clusters is chosen as *3*, which is evaluated empirically from cluster quality measure. The predictor attributes and number of years in different clusters as obtained by co-clustering method are shown in Table 1.

Table 1: Co-clustering results of years cluster with appropriate predictor attributes

Clusters	**Predictor attributes**	**Number of years**
Cluster 1	A1, A3, A5, A6, A7, A9, A11, A13, A14	33
Cluster 2	A4	12
Cluster 3	A2, A8, A10, A12, A15	20

Prediction of Indian Rainfall: Annual Indian summer monsoon rainfall (*ISMR*) is predicted using three models, namely, *RegTreeB*, *DecTreeB*, and *MultipleReg*. Models are built for each obtained cluster of years with its specific predictor set separately. Rainfall is expressed in terms of long period average of rainfall (*LPA*). Twenty years from *1993* to *2012* is considered as test-period. Final forecast of *ISMR* is presented as ensemble of forecasts given by individual cluster's model. Mean absolute errors for individual cluster and ensemble models are elaborated in Table 2. Ensemble models surpass individual cluster's models. It is also noted that weighted ensemble models outrun the simple mean ensemble model. All ensemble mean absolute errors by *RegTreeB* and *DecTreeB* are found to be around 5.0%, which are appreciable for predicting complex phenomenon of monsoon. In addition, *EnsModel 4* which assigns weight according to model's efficiency in prediction over the past years performs best with mean absolute error of 5.0% for test period *1993-2012*. The inter-annual variability of predicted and actual rainfall in terms of *LPA* is shown in Figure 2. It is observed that predicted values are closer to actual rainfall values during the considered test-period.

Table 2: Mean absolute errors (%) in prediction of Indian monsoon

Models	**Clus1**	**Clus2**	**Clus3**	**Ens1**	**Ens2**	**Ens3**	**Ens4**
RegTreeB	5.9	5.9	6.7	**5.5**	**5.1**	**5.3**	***5.0***
DecTreeB	6.2	6.1	7.5	**5.7**	**5.3**	**5.5**	**5.6**
MultipleReg	10.3	6.9	10.4	**8.2**	**6.8**	**6.4**	**6.7**

Comparison with State of Art Method: Proposed co-clustering-based ensemble model is compared with existing Indian Meteorological Department's (*IMD*) 16-parameter power regression model [2] and 8 and 10-parameter models [3], which give root mean square errors of 10.8%, 7.6%, and 6.4%, respectively, for period *1996-2002*. Proposed model *Ens4* with weighted ensemble of cluster models' prediction gives least root mean square error of *5.4%*. Other proposed models, namely, *Ens1*, *Ens2*, and *Ens3* give root mean square errors of *6.2%*, *5.8%*, and *6.1%*, respectively, outperforming all the three past *IMD* models(shown in Figure 3).

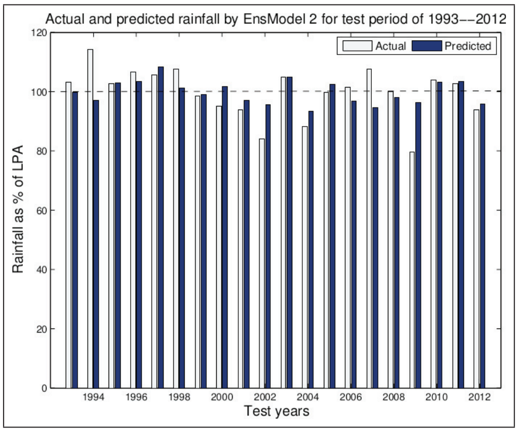

Figure 2: Actual and predicted rainfall as % of *LPA* by *Ens2* of *RegTreeB* model

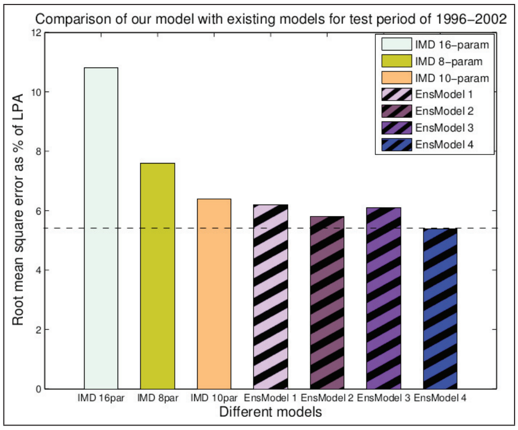

Figure 3: Comparison of proposed ensemble models with *IMD's* existing models [2, 3]

Comparison of Proposed Approach against Non-Clustered Approach: Proposed clustering-based ensemble models are compared with– (i) non-clustered, non-ensemble model with all fifteen predictors and all years of study, (ii) non-clustered, non-ensemble model with reduced clustered predictors obtained by co-clustering approach and all years of study. The result of comparison are shown in Table 3. The results clearly evince an improvement in prediction by proposed co-clustering-based ensemble approach over non-clustered general method.

Table 3: Comparison of proposed co-clustering-based ensemble models (Clus_Ens) against non-clustering methods (Non_Clus) (mean absolute errors(%))

Conditions	**Prediction Models**		
	RegTreeB	**DecTreeB**	**MultipleReg**
Error (Non_Clus) with all attributes	7.1	8.9	28.7
Error (Non_Clus) for clusters' attributes	7.6	8.0	8.4
Error of Ens1 (Clus_Ens)	**5.5**	**5.7**	**8.2**
Error of Ens2 (Clus_Ens)	**5.1**	***5.3***	**6.8**
Error of Ens3 (Clus_Ens)	**5.3**	**5.5**	***6.4***
Error of Ens4 (Clus_Ens)	***5.0***	**5.6**	**6.7**

5 Conclusions

Indian monsoon is a dynamic and complex phenomenon. India being an agricultural land, prediction of monsoon is very important for economic stability of the country. High variability in the monsoon process, makes the prediction task more challenging. It is highly difficult for a single model to frame such non-linearity. The article addresses this task by dual clustering of years and predictor parameters using co-clustering. Co-clustering approach clusters the years into groups with appropriate set of predictors for each groups. Three different models, namely, *RegTreeB*, *DecTreeB*, and *MultipleReg* are utilized for framing each clusters exclusively. Final forecast of monsoon is given by weighted ensemble of prediction given by cluster models. Ensemble model gives an mean absolute error of 5.0%, which is quite appreciable task for prediction of complex monsoon process. Finally, the prediction skill of proposed method is compared with existing *16*-parameter power regression model [2], and *8*-parameter, *10*-parameter models [3] used by *IMD*. The proposed approach gives root mean square error of 5.4%, which outperforms all three existing *IMD* models. The improvement of the clustering-based ensemble approach over non-clustered general approach is evident. Ensemble *MultipleReg* model shows an improvement of 20% in prediction over the non-clustered approach.

It is inferred that the proposed clustering and multi-model ensemble help in improvement in accuracy of prediction of monsoon process. Our future directions include identification of more global climatic predictors influencing Indian monsoon and their inclusion in prediction model. Also, other non-linear models are lined to be explored for even better prediction of Indian monsoon.

References

[1] I. S. Dhillon. Co-clustering documents and words using bipartite spectral graph partitioning. In *Proceedings of the seventh ACM SIGKDD International conference on Knowledge discovery and data mining*, pages 269–274. ACM, 2001.

[2] V. Gowariker, V. Thapliyal, S.M. Kulshrestha, G.S. Mandal, N. Sen Roy, and D.R. Sikka. A power regression model for long range forecast of southwest monsoon rainfall over India. *Mausam*, 42(2):125–130, 1991.

[3] M. Rajeevan, D.S. Pai, S.K. Dikshit, and R.R. Kelkar. IMD's new operational models for long-range forecast of southwest monsoon rainfall over India and their verification for 2003. *Current Science*, 86(3):422–431, 2004.

[4] M. Rajeevan, D.S. Pai, R. A. Kumar, and B. Lal. New statistical models for long-range forecasting of southwest monsoon rainfall over India. *Climate Dynamics*, 28(7-8):813–828, 2007.