Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 54 (2015) 271 – 280

Procedia Computer Science

Eleventh International Multi-Conference on Information Processing-2015 (IMCIP-2015)

Closed Item-Set Mining for Prediction of Indian Summer Monsoon Rainfall A Data Mining Model with Land and Ocean Variables as Predictors

H. Vathsala[a,*] and Shashidhar G. Koolagudi[b]

[a] *CDAC, Bengaluru 560 001, Karnataka, India*
[b] *National Institute of Technology, Karnataka, Mangalore 575 025, Karnataka, India*

Abstract

Practical application of data mining in scientific and engineering domains, when explored, pose many problems and provide interesting results. In this paper, we attempt to mine out association rules from 37 (1969–2005) years of Indian summer monsoon rainfall data and try its applicability in helping better prediction of Indian summer monsoon rainfall. We shortlist 36 variables as possible predictors of Indian summer monsoon rainfall based on previous literature and compare prediction using all 36 variables and prediction by selected attributes from derived association rules. Results show better performance in prediction of All India region, West central region and Peninsular region rainfall when attributes selection is employed as compared to all 36 variables used for prediction.

Peer-review under responsibility of organizing committee of the Eleventh International Multi-Conference on Information Processing-2015 (IMCIP-2015)

Keywords: Association rule; Attribute selection; CHARM; C4.5; ISMR prediction.

1. Introduction

Indian Summer Monsoon Rainfall (ISMR) is a complex phenomenon involving atmosphere, land, ocean and many other domains. When we speak about a large phenomenon like monsoon, we are aware that there are many variables involved that may very strongly relate to monsoon, variables that may relate weakly, variables that may not be related to monsoon but are present in climate scenario like in theories such as sea breeze[1]. A question arises here as to what are the variables to be considered for predicting or formulating association rules so as to give better results.

When there are many predictors, relationships between these predictors and rainfall is not well defined. For good accuracy in prediction, a few best predictors have to be picked based on some technique. Some techniques available to shortlist best predictors are step wise regression[2], a cross validation scheme suggested by[3]. Association rule mining[4] has found one of many applications in the field of feature selection. Feature selection based on association rule mining gets the feel of the considered data first by finding correlations between features involved so as to help us select only those features that are closely related to the subject of study. This assures a better prediction capability in good performance bracket.

*Corresponding author. Tel.: +91-080-2286-3100.
E-mail address: vathsala.h@gmail.com

Peer-review under responsibility of organizing committee of the Eleventh International Multi-Conference on Information Processing-2015 (IMCIP-2015)
doi:10.1016/j.procs.2015.06.032

The contribution of this study includes shortlisting possible ISMR predictors based on literature study, about 36 climate variables are shortlisted as possible ISMR predictors. Employing association rule mining for attribute selection in order to narrow down predictors that are closely related to different rainfall regions under study. Association rule mining based attribute selection also serves as a dimensionality reduction mechanism expected to overcome memory and time constraint imposed by desktop computers up to some extent. C4.5 classification algorithm based on decision tree is used as prediction tool. The combination of above mentioned techniques has shown a considerable improvement in the prediction results compared to results in previous study.

This paper is organised further with the literature review and data related information in section 2, Table 1. Gives a consolidated view of shortlisted variables in accordance to previous literature in the field. Section 3 describes deriving frequent item sets and association rules for attribute selection. Section 4 gives results and discussion. Conclusion is covered in section 5.

2. Literature Review, Data, Data Sources and Data Preparation

Long-range forecasting of all India monsoon rainfall started with the establishment of India Meteorological Department (IMD) in 1875. H. F. Blanford was the first to attempt a forecast of monsoon based on varying extent and thickness of the Himalayan snows. Since 1875 empirical and statistical studies on Indian monsoon rainfall have discovered many variables as predictors. Table 1 summarises a list of references of studies carried out on different variables related to ISMR. There have been constant efforts in the area of developing new models for predicting Indian summer monsoon rainfall. The efforts of[3,5–7], are noteworthy in recent times, But, in spite of all these efforts, it is believed that the achieved success is still not adequate and the scope for finding new prediction models and predictors are still wide open.

Selection of attributes or variables is a tough job for data mining community. This is due to the fact that data mining community may not have domain knowledge of the particular field under study. Such being the case, we go by research already done in ISMR. The idea is to find and pick out variables that relate to ISMR from papers already published, hereafter called the Large set and amongst large set variables, pick out a few strongly related variables for each region, hereafter called the selected attributes by applying association rule mining.

Table 1 lists all ISMR related variables with the paper references and provides an idea of related work in the considered area. The table contains a total of 15 entries with the first column *Paper reference* giving an idea of research work carried out on respective variable mentioned in the third column called *Variable*. The fourth column says weather the specified variable in the third column is included in the current work or not. The *Variable* column of the table includes those variables that are studied by domain experts as most likely predictors of ISMR and the *Considered / Not considered* column of the table specifies the Large set variables. The Last column namely *Total variables* gives a count of how many related variables of each category specified by *Variable* column are included in the study.

In finding Large set, we consider those variables that have been studied deeply with relatively more references. Predictors with less reference may be at infancy of research or may have been related to ISMR only for shorter period in time. Hence, we consider them not so predominant in predicting ISMR. Though April 500-mb ridge along 75 deg East, Eurasian Snow Cover and Snow Depth have many references, they are not considered in this study. Because, it requires domain expert to find the ridge on the mapped data and to identify snow cover from satellite images. No readymade information of these variables is available for the years considered in this study.

1) DSLP - DSLP is mean sea level pressure between grid point 130.0°E, 12.5°S[38]. According to[8]. DSLP tendency has good ISMR predicting capacity and is calculated as March-April-May average minus December-January-February average. Source: http://www.cpc.ncep.noaa.gov/data/indices/darwin.
2) El Nio events, ENSO SOI - NINO3 - As mentioned by[22], we calculate NINO3 by using monthly gridded SST anomalies averaged over NINO3 region and then averaged over the summer monsoon season, June, July, August and September (JJAS).

   NINO4 - According to[40], NINO4 area SST March-April-May average minus December-January-February average has a good correlation with ISMR.

   NINO3.4 - NINO3.4 calculations are done based on[41], as average of March-April-May Pacific surface temperature over NINO3.4 area.

Table 1. List of considered and not considered ISMR related variables with paper references.

SI no.	Paper reference	Variable	Considered / not considered	Total variables
1	5, 8–10	Darwin sea level pressure (DSLP)	Considered	1
2	9, 11–13	April 500-mb ridge along 75 deg East	Not considered	
3	10, 14–22	El Nio events, southern oscillation (ENSO SOI)	Considered	4
4	23	Negative Out going long wave radiation (OLR) anomalies	Not considered	
5	18–22	Equatorial Indian Ocean Oscillation (EQUINOO), Indian ocean sea surface temperature (SST), Equatorial Wind (EQWIN)	Considered	2
6	24–26	Eurasian Snow Cover, Snow Depth	Not considered	
7	5, 27–30	March-April-may minimum, maximum, Average air temperature Jodhpur, Ahmedabad, Bombay, Indore, Sagar, Akola	Considered	$16 * 3 = 18$
8	31	OLR anomalies over North Atlantic Ocean	Not considered	
9	5	East Coast of India Minimum Temperature (March)	Not considered	
10	5	North West Europe Mean Temperature (January)	Not considered	
11	5	North West India Pressure Anomaly (MAM)	Not considered	
12	5	Northern Hemisphere Pressure Anomaly (Jan April)	Not considered	
13	10, 28, 32–35	March-April-may Mean sea level pressure (MSLP) - Average mean sea level pressure Jodhpur, Ahmedabad, Bombay, Indore, Sagar, Akola, MSLP over india	Considered	6
14	36–38	200 hPa meridional component of wind of May at Bombay, Delhi, Madras, Nagpur and Srinagar	Considered	5
15	28	Wind on surface and 500hpa	Not considered	
				Total = 36

ENSO-SOI - (standard Tahiti sea level press - standard Darwin sea level press) is prepared as predictor in accordance to[42]. Source: http://www.cpc.ncep.noaa.gov/data/indices/soi.

E1 NINO Index SST - NINO3, NINO4 and NINO3.4.
Source: http://www.cpc.ncep.noaa.gov/data/indices/sstoi.indices.

3) EQUINOO - EQUINOO is the oscillation of Indian Ocean with enhanced convection over the West Equatorial Indian Ocean (WEIO) (50° to 70°E, 10°S to 10°N)[43] and reduced convection over East Equatorial Indian Ocean (EEIO) (90° to 110°E, 10°S to EQ)[43] (positive phase) and another with anomalies of the opposite signs (negative phase). It has indices based on Indian ocean SST and the zonal component of the surface wind at the equator (60°E to 90°E, 2.5°S to 2.5°N) called EQWIN. Prepared readymade data is available in[44] and we directly pick them without making any changes.

   Indian ocean SST. Source:[44]provided as EQUINOO phase till 2005.

   Zonal component of the surface wind EQWIN Source:[44]provided as EQUINOO Index till 2005.

4) March-April-May minimum (min), maximum (max), Average air temperature Jodhpur, Ahmedabad, Bombay, Indore, Sagar and Akola - Observed data as provided in IITM website has $1° \times 1°$ longititude and latitude resolution of averaged March-April-May values for max, min and average temperatures.
   Source: http://cccr.tropmet.res.in/cccr/getUI.do?dsid=id-a4ea271c42&varid=winter_mean_temp-id-a4ea271c42&auto=true.

5) MSLP - Average Mean Sea Level Pressure Jodhpur, Ahmedabad, Bombay, Indore, Sagar and Akola - Observed data downloaded from NCEP-NCAR Reanalysis is in $2.5° \times 2.5°$ longititude and latitude resolution, it is interpolate and $0.5° \times 0.5°$ resolution grid data is used.
   Source: http://iridl.ldeo.columbia.edu/maproom/Global/Atm_Circulation/Sea_Level_Pres.html.

6) 200 hPa meridional component of wind for May at Bombay, Delhi, Madras, Nagpur and Srinagar - Observed data downloaded from NCEP-NCAR Reanalysis is in $2.5° \times 2.5°$ longititude and latitude resolution, it is interpolate and $0.5° \times 0.5°$ resolution grid data is used.
   Source: http://iridl.ldeo.columbia.edu/maproom/Global/Atm_Circulation/Wind_SST.html.

7) Rainfall Data Homogenous region wise rainfall data for All India region, West central region and Peninsular region. Source: ftp://www.tropmet.res.in/pub/data/rain/iitm-regionrf.txt.

2.1 Data representations and ranging

In order to use the available data for data mining, we club similar instances into one range and these ranges are given names. Data are divided into 5 ranges, the threshold values are determined by identifying the values at ± 1 and ± 0.5 Standard Deviation (SD) from the average.

We name the ranges in a pattern - the name of the variable followed by the range number, for example DSLP has five ranges and is represented as DSLP-1, DSLP-2, DSLP-3, DSLP-4 and DSLP-5 respectively. Once we find association rules, we can reverse map these names to numbers for our understanding.

3. Deriving Frequent Item-Sets and Association Rules for Attribute Selection

Since the Large set variables are 36 in number, addition of rainfall variable to it gives a total of 37 variables. With such huge dimensionality, the task is to find an apt technique and algorithm that can generate frequent Item-sets from huge dimensional data.

The basic concept of association rule mining is to find the set of all item-sets that are frequently occurring in the database in the form of transactions and further these frequent item-sets are used to frame rules. Support and confidence determine how strong the association rule is

Definition 1. *Let $I = I1, I2, I3, In$ be a set of n attributes called items. Let $D = T1, T2, T3, \ldots, Tm$ be a set of m transactions where each transaction has a transaction ID T, such that each transaction contains a subset of items in I. An association rule is an implication of the form $X \Rightarrow Y$ such that $X, Y \subseteq I$ and $X \cap Y = \emptyset$. Where X is called the antecedent and Y is called the consequent.*

Support and Confidence are the two parameters to determine the interestingness of a rule. There are many such measures like lift, f-measure, cosine etc. to gauge the interestingness of the rule. Support and Confidence are well known in data mining community.

Definition 2. *Support is defined as the proportion of transactions that contains X.*

$$Support(X) = \frac{X.count}{m}$$

Definition 3. *Confidence of $X \Rightarrow Y$ is defined as the measure of how often items in Y appear in transaction that contains X.*

$$Confidence(X \Rightarrow Y) = \frac{(X \cup Y).count}{X.count}$$

The common example cited for association rule mining is market basket analysis, this being a prototypical example, where recorded sales data at a large grocery or a departmental stores with items represented by products are analysed. These types of databases are generally sparse, i.e. the longest frequent Item-sets are relatively short. However, there are many real world data-sets that are very dense.

When the specified frequency threshold requirement is low. It is seen that the set of frequent items and association rules can rapidly grow to large numbers. The larger the number of frequent items, the more the number of rules derived. Many of these rules are redundant rules, thus becoming a constraint in terms of memory and processing speed requirement. Dense item-sets usually result in exponential number of frequent item-sets. This is true even in sparse data-sets but can be manageable many of the times. In dense data-sets it becomes difficult to derive frequent item-sets itself, leave alone generating association rules from frequent item-sets.

Techniques used to overcome large frequent Item-set problems without leaving out important associations and compromising on the quality are being explored. Some of these techniques are to mine maximal frequent item-sets and closed frequent item-sets.

Maximal Frequent item-set was invented based on the requirement of considering only the frequent item-set that has the maximum number of items bypassing all the sub-item-sets. Item-set is maximal frequent if none of its immediate supersets is frequent.The drawback of maximal frequent item-set is that it may leave out some frequent item-sets

that may result in important association rules, because it concentrates only on frequency and not their support. This problem is overcome by closed set. An item-set is closed if none of its immediate supersets has the same support as this item-set has. Finding closed frequent item-sets can be of great help to prune item-sets that are not needed and to find the right association rules. In 1999, researchers came up with an article (Discovering Frequent Closed Item-sets for association Rules by Nicolas Pasquier, Yves Bastide, Rafik Taouil, and Lotfi Lakhal). They proposed to mine only closed frequent item-sets using an algorithm called A-CLOSE. This was followed by many other algorithms like CHARM, CLOSET etc. based on closed frequent item-sets,[45] have proved that, it outperforms the previous methods by an order of magnitude or more and it is also linearly scalable in the number of transactions and the number of closed item-sets found.

The current research data considered is dense (36 variables). Drought, deficit, excess and flood occur in low frequencies. Hence, mining frequent item-set requires setting minimum support to very low values, thus resulting in a big deal of memory and speed requirement. For example, in the proposed work, we consider 37 (1969 to 2005) -1 (leave out 1 cross validation) $=$ 36 years of data for association rule mining. On examining the data it is seen that excess rainfall in All India region occurs only 4 times in 36 years. As a result, to find frequent item-set of frequency at least 2, according to the formula of support specified in Definition 2 the minimum support, has to be set to 0.058, thus resulting in millions of frequent item-sets.

To overcome this problem of generating enormous amount of unwanted frequent item-sets, we propose to mine closed item-sets using CHARM[46] algorithm. Thus resulting in approximately around two thousand closed frequent item-sets for the example mentioned. We set the minimum support to 0.058 and select top 2 to 4 association rules ranked according to confidence in each of the rainfall ranges. The implementation of CHARM available in sequential pattern mining frame work (SPMF)[49] is used in the current research.

4. Classification Algorithm for Prediction

We use classification algorithm for constructing a prediction model. Decision tree concept of classification algorithm is employed by making the algorithm train itself by creating a decision tree through the training data-set. The rules generated by employing the decision tree constructed can further be used in classifying the test data-set into known classes resulting in a prediction model. In Current work C4.5[47], a classification algorithm is used as the tool of prediction. This algorithm is one of the efficient algorithms which appeared in[48]. C4.5 is a widely used free data mining tool and is a descendent of an earlier system called ID3[47].

The prediction model consists of decision tree classification algorithm with its options verbosity level set to 3, force subsetting set to true and minimum objects set to 6. The options were set to avoid over fitting and were arrived at after a number of trials providing satisfactory results. The problem with decision tree is that decision tree learners create biased trees if some classes dominate. It is therefore recommended to balance the data-set prior to fitting with the decision tree. The variables selected from top Association rules ranked according to confidence precisely are only those variables that have balanced data, hence giving the decision tree learner, balanced data as input. Using the results of this prediction model, we make a comparison between prediction Of ISMR considering the large set variables and selected attributes from section 3.0. Association rules and selected attributes of each rainfall region considered are presented in Tables 2, 3, 4.

4.1 Results and discussion

Prediction through decision tree algorithm C4.5 has given the above results. Table 5 presents Standard Deviation (SD), Standard Deviation Error (SDErr) and Root Mean Square Error (RMSE) calculated against observed data ranges. We have worked with 37 (1969–2005) years of data for ranging and have employed leave out one cross validation scheme, accordingly, for rainfall prediction, association rules are derived from data of all years other than the data of the year to be predicted. Similarly, classification algorithm is trained using data of all years other than the data of the year to be predicted. The results show that the attribute selection based technique works good overall this is demonstrated by the SD, SDErr and RMSE of each region with considerable improvement when selected attributes are used for prediction as compared to large set attributes. Figure 1 of All India rainfall prediction, Fig. 2 of West central

Table 2. Lists association rules their confidence and selected attributes for all India region.

All India Rainfall	Rule	Confidence	Selected attributes
Drought	Eqwin-3 NINO4-5 Equinoo-phase-2 ⇒ Allindia-rainfall-1	0.8	Eqwin,
	DSLP-3 SOI-3 ⇒ Allindia-rainfall-1	0.8	NINO4,
			Equinoo-phas,
Deficit	Sr-wnd-3 Bb-Maxtmp-3 NINO4-3 ⇒ Allindia-rainfall-2	0.8	DSLP,
			SOI,
			Sr-wnd,
Normal	Jd-mslp-3 DSLP-4 ⇒ Allindia-rainfall-3	0.8	Bb-Maxtmp,
	Bb-mslp-2 Ak-mslp-2 Id-mslp-2 ⇒ Allindia-rainfall-3	0.8	Jd-mslp,
	Id-Mintmp-3 Sa-Mintmp-3 Ng-wnd-2 ⇒ Allindia-rainfall-3	0.8	Bb-mslp,
Excess	Equinoo-phase-2 NINO3.4-3 Ma-wnd-2 ⇒ Allindia-rainfall-4	0.8	Ak-mslp,
	Jd-Mintmp-3 Id-mslp-2 Jd-mslp-2 ⇒ Allindia-rainfall-4	0.8	Id-mslp,
	Jd-Mintmp-3 Jd-Maxtmp-3 Jd-Meantmp ⇒ Allindia-rainfall-4	0.8	Id-Mintmp,
			Sa-Mintmp,
			Ng-wnd,
			NINO3.4,
Flood	Ak-Maxtmp-3 Equinoo-phase-2 Ak-Meantmp-3 ⇒ Allindia-rainfall-5	0.44	Ma-wnd,
			Jd-Maxtmp,
			Jd-Mintmp,
			jd-Meantmp,
			Ak-Maxtmp,
			Ak-Meantmp

Table 3. Lists association rules their confidence and selected attributes for west central region.

West central Rainfall	Rule	Confidence	Selected attributes
Drought	Ah-Maxtmp-3 Eqwin-3 Id-Maxtmp-3 Euinoo-phase-2 Ah-Meantmp-3 ⇒ Westcentral-rainfall-1	0.8	Bb-wnd,
	Bb-wnd-5 ⇒ Westcentral-rainfall-1	0.63	Ah-Maxtmp,
			Eqwin,
			Id-Maxtmp,
Deficit	DSLP-3 SOI-3 ⇒ Westcentral-rainfall-2	0.8	Euinoo-phase,
	Sr-wnd-3 Bb-Maxtmp-3 NINO4-3 ⇒ Westcentral-rainfall-2	0.8	Ah-Meantmp,
			DSLP,
			SOI,
			Sr-wnd,
Normal	Bb-wnd-2 Ma-wnd-2 ⇒ Westcentral-rainfall-3	0.63	Bb-Maxtmp,
	Ak-Maxtmp-2 Ak-mintmp-2 Ak-Meantmp-2 ⇒ Westcentral-rainfall-3	0.63	NINO4,
	Sa-Meantmp-3 Id-Mintmp-3 Sa-Mintmp-3 ⇒ Westcentral-rainfall-3	0.63	Ma-wnd,
			Ak-Maxtmp,
			Ak-Mintmp,
			Ak-Meantmp,
Excess			Sa-Meantmp,
			Id-Mintmp,
			Sa-Mintmp,
Flood	DSLP-1 Euinoo-phase-2 ⇒ Westcentral-rainfall-4	0.8	NINO3.4,
	Bb-wnd-3 NINO4-3 Euinoo-phase-2 ⇒ Westcentral-rainfall-4	0.8	Jd-mslp
	Euinoo-phase-2 NINO3.4-3 SOI-3 Jd-mslp-2 ⇒ Westcentral-rainfall-4	0.8	

India rainfall prediction with selected attributes gives considerably better results with improved SD, SDErr and RMSE as shown in Table 5. Figure 3 of Peninsular region also shows improved results when selected attributes are used but is not as good as the previous two regions. This may be partially due to the absence of selected attributes for excess and drought in peninsular region.

If we consider the prediction model as such, the model performs quite well in All India, West central and Peninsular region. This is evident when we see the RMSE of this region. Peninsular India does not show a considerable change

Table 4. Lists association rules their confidence and selected attributes for west central region.

West central Rainfall	Rule	Confidence	Selected attributes
Drought			DSLP, SOI, Sr-wnd, Ng-wnd,
Deficit	DSLP-3 SOI-3 ⇒ Peninsular-rainfall-2	0.8	NINO4,
	Sr-wnd-3 Ng-wnd-2 ⇒ Peninsular-rainfall-2	0.63	Ma-wnd,
	NINO4-3 Ma-wnd-3 ⇒ Peninsular-rainfall-2	0.63	Euinoo-phase, Bb-wnd,
Normal	NINO4-3 Euinoo-phase-2 Ng-wnd-2 ⇒ Peninsular-rainfall-3	0.789	Id-Maxtmp,
	Bb-wnd-2 Bb-Maxtmp-3 Id-Maxtmp-3 ⇒ Peninsular-rainfall-3	0.789	Ak-Maxtmp,
	Ak-Maxtmp-3 NINO4-3 NINO3.4-3 ⇒ Peninsular-rainfall-3	0.789	NINO3.4,
	NINO3-3 NINO4-3 NINO3.4-3 ⇒ Peninsular-rainfall-3	0.789	NINO3, Ak-Mintmp
Excess			
Flood	Ak-Mintmp-2 Euinoo-phase-2 ⇒ Peninsular-rainfall-3	0.63	

Table 5. Performance of the model in terms of standard deviation, standard deviation error, root mean square error between the observed ranges and predicted ranges of summer monsoon rainfall.

	Large set All India	Selected All India	Large set West central	Selected West central	Large set Peninsular	Selected Peninsular
Standard Deviation (Current work)	0.8333333	0.745355	1.0671873	0.9428090	1.1785113	1.1180339
Standard Deviation Error (Current work)	0.1369991	0.122535	0.1754445	0.1549968	0.1937460	0.1838036
Root Mean Square error in range (Current work)	0.8219949	0.735214	1.0654272	0.92998111	1.1624763	1.1028219
Root Mean Square error as per [21]		1.02		0.99		0.71

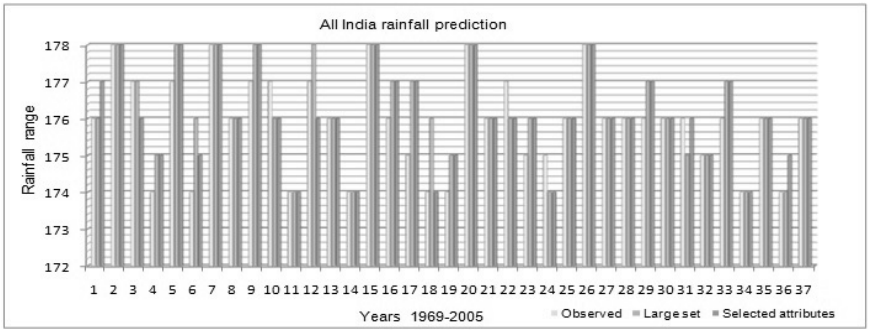

Fig. 1. Plot of all India region rainfall.

with selected attributes. The most likely reason for this might be due to the fact that ISMR is not the chief rainy season for peninsular India[7]. Hence, these predictors may not be the correct predictors for rainfall prediction in peninsular region. A combination of ISMR predictors and North-East monsoon predictors can be experimented upon for better understanding rainfall in peninsular region. The RMSE of the current work has shown improvement over previous

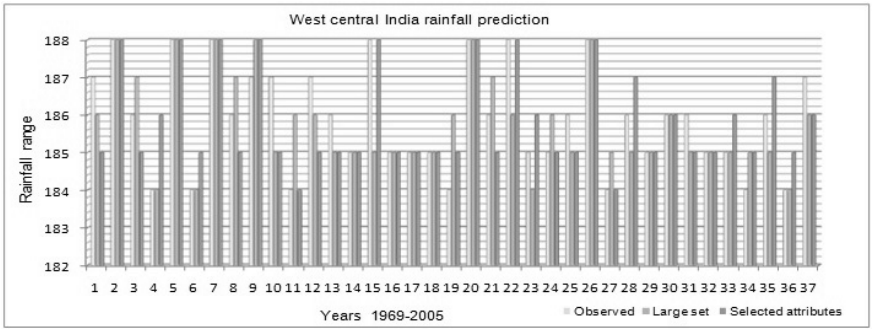

Fig. 2. Plot of West central region rainfall.

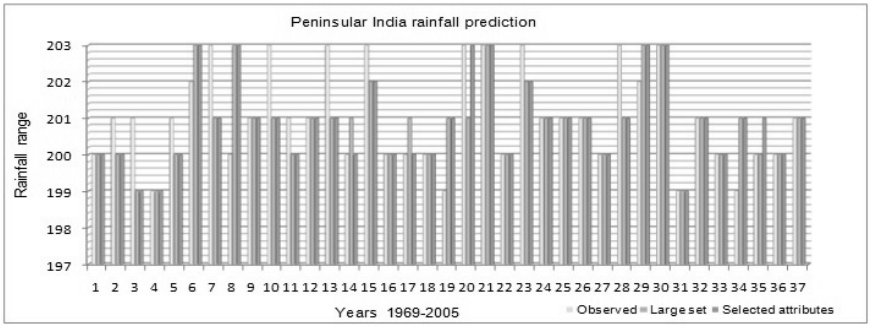

Fig. 3. Plot of Peninsular region rainfall.

research, the last row in Table 5 shows the RMSE of applying Fuzzy association rule mining with ENSO index and EQWIN index as predictors. In comparison with results of[6,21], in current work All India region and West central region show improved prediction. This might be because of the inclusion of many other climate indices related to ISMR as suggested in[6,21], peninsular region has shown increased RMSE, there might be reasons for this, first might be due to the addition of more climate indices related to summer monsoon, were as summer monsoon season is not chief rainy season for Peninsular region, Peninsular region gets major share of its rains from North-East monsoon. Second might be that the number of years considered in the study may not be sufficient to bring out the hidden relationship between the climate variables. As a result, we could not derive association rules for excess rainfall and drought in peninsular region, this might have resulted in increased RMSE.

5. Summary and Conclusion

Possible predictors of ISMR, about 36 variables, are filtered from a survey, without considering the time scale in which these variables were active predictors. Application of association rule mining for attribute selection from 36 variables for predicting ISMR is experimented and considerably good results are obtained for All India, West central regions. Peninsular region also exhibits improved results using the attribute selection technique but does not show a considerable improvement. A practical problem of applying data mining techniques on dense data-sets with respect to large frequent item-sets due to low minimum support is discussed. Closed item-sets using CHARM algorithm are mined and used for attribute selection. Here, closed item-sets are used instead of all frequent item-sets to overcome the memory and speed requirement constraint without compromising in obtaining useful association rules. Classification algorithm C4.5 is used as a prediction tool. The model performs well in majority of the regions.

Further research directions - ISMR being a complex phenomenon, addition of indices like Eurasian snow cover, Eurasian snow depth, April 500-mb ridge along 75 deg East etc., considering longer interval of data, by increasing the

number of years in research data, instead of using crisp partitioning employing fuzzy partitioning and other prediction models like neural networks with attribute selection instead of C4.5 classification algorithm can be experimented to improve the results.

References

[1] D. Bala Subrahamanyam, K. Sen Gupta, Sudha Ravindran and Praveena Krishnan, Study of Sea Breeze and Land Breeze Along the West Coast of Indian Sub-Continent Over the Latitude Range 15°N to 8°N During INDOEX IFP-99 (SK-141) Cruise, Current Science (Supplement), vol. 80, 10 April (2001).

[2] D. S. Wilks, Statistical Methods in the Atmospheric Sciences (An Introduction), *International Geophysics Series*, Academic Press, San Diego, CA, USA, pp. 467, (1995).

[3] T. DelSole and J. Shukla, Linear Prediction of Indian Monsoon Rainfall, Engineering Applications using MATLAB, Prof. Emilson Pereira Leite (Ed.), ISBN: 978-953-307-659-1, InTech, (2001).

[4] R. Agrawal, T. Imieliåski and A. Swami, Mining Association Rules Between Sets of Items in Large Databases Proceeding SIGMOD'93, *International Conference on Management of Data*, ACM New York, NY, USA 1993, pp. 207–216, (1993), ISBN:0-89791-592-5 doi:10.1145/170035.170072

[5] P. Guhathakurta, M. Rajeevan and V. Thapliyal, Long Range Forecasting Indian Summer Monsoon Rainfall by a Hybrid Principal Component Neural Network Model, *Atmos. Phys.*, vol. 71, pp. 255–266, Springer-Verlag, (1999).

[6] C. T. Dhanya and D. Nagesh Kumar, Data Mining for Evolution of Association Rules for Droughts and Floods in India using Climate Inputs, *Journal of Geophysical Research*, vol. 114, D02102, (2009), doi:10.1029/2008JD010485.

[7] S. Satishkumar, Kashid and Rajib Maity, Prediction of Monthly Rainfall on Homogeneous Monsoon Regions of India Based on Large Scale Circulation Patterns using Genetic Programming, *Journal of Hydrology*, pp. 454–455, 26–41, (2012).

[8] J. Shukla and A. Daniel Paolino, The Southern Oscillation and Long-Range Forecasting of Summer Monsoon Rainfall Over India, Monthly Weather Review, *American Meterological Society*, vol. 111, (1983).

[9] J. Shukla and D. A. Mooley, Empirical Prediction of the Summer Monsoon Rainfall over India, *Mon. Wea. Rev.*, vol. 115, pp. 695–704, (1987).

[10] K. Krishna Kumar and R. Kleeman, Epochal Changes in Indian Monsoon ENSO Precursors, *Geophysical Research Letters*, vol. 26, pp. 75–78, (1999).

[11] D. A. Mooley, B. Parthasarathy and G. B. Pant, Relationship between Indian Summer Monsoon Rainfall and Location of the Ridge at the 500-mb Level along 75°E, *J. Climate Appl. Meteor.*, vol. 25, pp. 633–640, (1986).

[12] A. K. Banerjee, P. N. Sen and C. R. V. Raman, On Foreshadowing Southwest Monsoon Rainfall over India with Midtropospheric Circulation Anomaly of April, *Indian J. Met. Hydrol. Geophys.*, vol. 29, pp. 425–431, (1978).

[13] K. Krishna Kumar, K. Rupa Kumar and G. B. Pant, Pre Monsoon Ridge Location over India and its Relation to Monsoon Rainfall, *J. Climate*, vol. 5, pp. 979–986, (1992).

[14] Rasmusson, M. Eugene and H. Thomas Carpenter, The Relationship between Eastern Equatorial Pacific Sea Surface Temperatures and Rainfall over India and Sri Lanka, *Mon. Wea. Rev.*, vol. 111, pp. 517–528, (1983).

[15] C. F. Ropelewski and M. S. Halpert, Precipitation Patterns Associated with the High Index Phase of the Southern Oscillation, *J. Climate*, vol. 2, pp. 268–284, (1989).

[16] Clark, Christina Oelfke, E. Julia Cole and J. Peter Webster, Indian Ocean SST and Indian Summer Rainfall: Predictive Relationships and their Decadal Variability, *J. Climate*, vol. 13, pp. 2503–2519, (2000).

[17] J. M. Slingo and H. Annamalai, The El Nio of the Century and the Response of the Indian Summer Monsoon, *Mon. Wea. Rev.*, vol. 128, pp. 1778–1797, (1997, 2000).

[18] S. Gadgil, The Indian Monsoon and its Variability, *Annual Review of Earth Planetary Science*, vol. 31, pp. 429–67, (2003).

[19] S. Gadgil, M. Rajeevan and R. Nanjunduah, Monsoon Prediction: Why Yet Another Failure, *Current Science*, vol. 88(9), pp. 1389–1400, (2005).

[20] Rajib Maity and D. Nagesh Kumar, Bayesian Dynamic Modeling for Monthly Indian Summer Monsoon Rainfall using El Nino–Southern Oscillation (ENSO) and Equatorial Indian Ocean Oscillation (EQUINOO), *JOU Journal of Geophysical Research*, vol. 111, (2006), D07104, doi:10.1029/2005JD006539.

[21] C. T. Dhanya and D. Nagesh Kumar, Data Mining for Evolving Fuzzy Association Rules for Predicting Monsoon Rainfall of India, *Journal of Intelligent Systems*, vol. 18, no. 3, (2009).

[22] Chie Ihara, Yochanan Kushnir and A. Mark Canea and H. Victor De La Penab, Indian Summer Monsoon Rainfall and its Link with ENSO and Indian Ocean Climate Indices, *International Journal of Climatology Int.*, vol. 27, pp. 179–187, (2006).

[23] Bin Wang and Zhen Fan, Choice of South Asian Summer Monsoon Indices, *Bulletin of the American Meteorological Society*, (1998).

[24] A. S. Bamzai and J. Shukla, Relation between Eurasian Snow Cover, Snow Depth and the Indian Summer Monsoon: An Observational Study, *J. Climate*, vol. 12, pp. 3117–3132, (1999).

[25] A. D. Vernekar, J. Zhou and J. Shukla, The Effect of Eurasian Snow Cover on the Indian Monsoon, *J. Climate*, vol. 8, pp. 248–266, (1995).

[26] B. Parthasarathy and S. Yang, Relationships between Regional Indian Summer Monsoon Rainfall and Eurasian Snow Cover, *Adv. Atmos. Sci.*, vol. 12, pp. 143–150, (1995).

[27] D. A. Mooley and D. A. Paolino, A Predictive Monsoon Signal in the Surface Level Thermal Field Over India, *Mon. Wea. Rev.*, vol. 116, pp. 256–265, (1988).

[28] K. Shashikanth and Subimal Ghosh, Fine Resolution Indian Summer Monsoon Rainfall Projection with Statistical Downscaling, *International Journal of Chemical, Environmental & Biological Sciences (IJCEBS)*, vol. 1, no. 4, ISSN 2320-4079; EISSN pp. 2320–4087, (2013).
[29] B. Parthasarathy, K. Rupa Kumar and N. A. Sontakke, Surface and Upper Air Temperatures over Indian in Relation to Monsoon Rainfall, *Theoretical and Applied Climatology*, vol. 42, pp. 93–110, Springer-Verlag, (1990).
[30] K. Bhaskara Rao, C. V. Naidu and O. S. R. U. Bhanukumar, Forecasting of Indian South-West Monsoon Rainfall, *Meteorological Applications*, vol. 8, no. 2, Article First Published Online: 29 December (2006).
[31] A. K. Srivastava, M. Rajeevan and R. Kulkarni, Teleconnection of OLR and SST Anomalies over Atlantic Ocean with Indian Summer Monsoon, *Geophysical Research Letters*, vol. 29, no. 8, pp. 125-1–125-4, (April 2002).
[32] B. Parthasarathi, K. Rupa Kumar and A. A. Munot, Evidence of Secular Variations in Indian Monsoon Rainfall-Cerculation Relationships, *Journal of Climate*, vol. 4, American Metorological Society, (1991).
[33] B. Parthasarathy, K. Rupa Kumar and A. A. Munot, Surface Pressure and Summer Monsoon Rainfall over India, *Advances in Atmospheric Sciences*, vol. 9, no. 3m, (August 1992), Springer.
[34] M. Lal, L. Bengtsson, U. Cubasch, M. Esch and U. Schlese, Synoptic Scale Disturbances of the Indian Summer Monsoon as Simulated in a High Resolution Climate Model, *Climate Research*, vol. 5, pp. 243–258, (1995).
[35] S. D. Patil, H. N. Singh, S. D. Bansod and Nityanand Singh, Trends in Extreme Mean Sea Level Pressure and Their Characteristics during the Summer Monsoon Season over the Indian Region, *International Journal of Remote Sensing*, vol. 32, no. 3, Special Issue: Remote Sensing and Climate Change, (2011).
[36] R. K. Verma and P. P. Kamte, Statistical Technique for Long-Range Forecasting of Summer Monsoon Activity over India, *Proceeding of Summer Monsoon Activity over India. Proceedings of Symposium on the Probabilistic and Statistical Methods in Weather Forecasting*, 8–12 September, Nice, WMO, Geneva, pp. 303–307, (1980).
[37] P. V. Joseph, R. K. Mukhopadhyaya, W. V. Dixit and D. V. Vaidya, Meridional Wind Index for Long Range Forecasting of Indian Summer Monsoon Rainfall, *Mausam*, vol. 32, pp. 31–34, (1981).
[38] B. Parthasarathy, K. Rupa Kumar and V. R. Deshpande, Indian Summer Monsoon Rainfall and 200 mb Meridional Wind Index: Application for Long Range Prediction, *Int. J. Climatol.*, vol. 11, pp. 165–176, (1991).
[39] Rasmusson, M. Eugene and H. Thomas Carpenter, Variations in Tropical Sea Surface Temperature and Surface Wind Fields Associated with the Southern Oscillation/El Nio, *Mon. Wea. Rev.*, vol. 110, pp. 354–384, (1982).
[40] K. Krishnakumar, M. K. Soman and K. Rupa Kumar, Seasonal Forecasting of Indian Summer Monsoon Rainfall: A Review Weather, vol. 150, pp. 449–467, (1995).
[41] N. A. Rayner, D. E. Parker, E. B. Horton, C. K. Folland, L. V. Alexander, D. P. Rowell, E. C. Kent and A. Kaplan, Global Analyses of SST, Sea Ice and Night Marine Air Temperature Since the Late Nineteenth Century, *J. Geophys. Res.*, vol. 108, pp. 4407, (2003), doi:10.1029/2002JD002670.
[42] S. C. Chakravarty, Sea Surface Temperature (SST) and the Indian Summer Monsoon, Scientific and Engineering Applications using MATLAB, *Prof. Emilson Pereira Leite (Ed.)*, (2011), ISBN: 978-953-307-659-1, InTech, Available from: http://www.intechopen.com/books/scientific-and-engineering-applications-using-matlab/seasurface-temperature-sst-and-the-indian-summer-monsoon.
[43] P. A. Francis and Sulochana Gadgil, A Note on New Indices for the Equatorial Indian Ocean Oscillation, *J. Earth Syst. Sci.*, vol. 122, no. 4, pp. 1005–1011, August (2013).
[44] B. V. Charlotte, Dhanya and Basil Mathew, EQUINOO: The Entity and Validity of this Oscillation to Indian Monsoon, RESEARCH INVENTY: *International Journal of Engineering and Science*, ISBN: 2319–6483, ISSN: 2278–4721, vol. 1, no. 11, pp. 45–54, December (2012).
[45] M. J. Zaki and C. J. Hsiao, CHARM: An Efficient Algorithm for Closed Association Rule Mining (Tech. Rep. No. 99–10), Rensselaer Polytechnic Institute, *Department of Computer Science*, (1999).
[46] M. J. Zaki and C. J. Hsiao, CHARM: An Efficient Algorithm for Closed Item-set Mining, *Proceedings of the 2002 SIAM International Conference on Data Mining*, pp. 457–473, (2002).
[47] J. R. Quinlan, Discovering Rules by Induction from Large Collections of Examples, In: Michie D (ed), Expert Systems in the Micro Electronic Age, Edinburgh University Press, Edinburgh Quinlan JR (1993) C4.5: Programs for Machine Learning, Morgan Kaufmann Publishers, San Mateo, (1979).
[48] XindongWu, Vipin Kumar, J. Ross Quinlan, Joydeep Ghosh, Qiang Yang, Hiroshi Motoda, J. Geoffrey McLachlan, Angus Ng, Bing Liu, S. Philip Yu, Zhi-Hua Zhou, Michael Steinbach, J. David Hand and Dan Steinberg, Top 10 Algorithms in Data Mining, *Knowl. Inf. Syst.*, vol. 14, pp. 1–37, (2008).
[49] P. Fournier-Viger, A. Gomariz, A. Soltani and T. Gueniche, SPMF: Open-Source Data Mining Platform, (2013), http://www.philippe-fournier-viger.com/spmf/