International Journal of Recent Technology and Engineering (IJRTE)
ISSN: 2277-3878, Volume-8 Issue-3, September 2019

An Experiment in Weather Prediction for North East and South West Monsoon Over Chennai

S.Meganathan, R.Bala Krishnan, A.Sumathi, S.Sheik Mohideen Shah, J.Senthilkumar

Abstract— There is a growing demand for spot specific forecast. Presently this has to be extracted from the regional forecast based on synoptic models. Synoptic models require input from various observatories of regions or the country and the central analysis centre is required for generating the synoptic charts. But recently the authors have established the potential of local data alone as a continuous time scale for use in effective local forecast using data mining techniques. Following the same association rule mining and classifier approach is tried for the forecast of wet and dog days on North East Monsoon and South West Monsoon months for the Chennai region with Latitude 13°11' N and Longitude 80°11' E, a coastal station over Bay of Bengal in South India and results are presented.

Keywords Association rule mining, classifier approach, rainfall forecast, data mining.

I. INTRODUCTION

Conventionally synoptic charts are used for weather forecasting. For 'spot forecast' or 'stations specific forecast' whose need is increasing nowadays due to the increase in metropolis, industries and the allied activities, the same has to be extracted from the charts. Meteorologists have to use their knowledge of identification and motion of synoptic systems and the associated regions of confluence, divergence etc in this regard. But the practical experience is that the forecast has a tendency to go off the mark. Paucity of data in charts, sudden slowing of synoptic system etc., may be some of the reasons. Radar observations and the development of station climatology are helpful to some extent in this regard. Isolated attempts using the methodology (Sivaramakrishnan, 1983, 1988; Mohanty, 1994; Seetharam, 2009) [5,7] have been made. Nevertheless with the advent of computers more efficient techniques are possible.

The terminology "Data mining" is a popular mechanism in order to knob this kind of peculiar needs. It finds applications on many fields [1,2] like business astuteness, fraudulent activity detection, credit and market basket analysis, computational biotechnology. Recently the authors Sivaramakrishnan [9]presented an approach to obtain association rules (ASS-Rule), which could be utilized for prediction process over the weather data and in convinced case studies of the authors Sivaramakrishnan and Meganathan, 2012a, 2012b, 2013, 2014 [10,11,12,13] the obtained consequences were encouraging.

The Chennai region data with Latitude value 13°11' N and Longitude value 80°11' E is a metropolitan city of peninsular India along the east coast. Apart from many industries, a variety of vital installations are there and a lot of social activities and sports take place. All these things demand day specific forecast regarding rain and human comfort needs the lowest minimum and highest maximum temperatures during winter and summer respectively. This station gets copious rainfall during October, November and December of Northeast (NE) monsoon [4]. Even during Southwest (SW) monsoon months of June to September there are days of thundershowers [6,8].The forecast for wet days and dry days in both season [14] will be useful. Also the prediction of hot and very cold days throughout the subsequent seasons is significant for thermal level ease for human beings.

II. METHODOLOGIES AND RAW INFORMATION

The region Chennai with Latitude 13°11' N and Longitude 80°11' E is an inshore location of South India. India meteorological department maintains a weather observatory since long. The comprehensive précis of the exterior information for the duration 1961 to 2010 is extracted from the data repository of World Meteorological Organization located at, Asheville, USA. The months October, November and December are deliberated for North-East monsoon period and the moths June, July, August and September are deliberated for South-West monsoon. The synaptic parameters like Temperature (TP), Saturation level or Dew-Point (DP), Speed value of Wind (WS), Visibility (VISB) and Rainfall-Precipitation (RP) were debated for examination. The data sets which are extracted contain the widespread atmospheric berth of 24 and 48 hours earlier than the actual incidence of the weather event.

The DM practices such as cleaning, content selection, information transformation are followed at this phase. The atmospheric content is pre-processed and transformed into mining attributes by enforcing discretization procedure with un-supervised learning mechanism [3].

Manuscript published on 30 September 2019
* Correspondence Author
S.Meganathan*, Faculty of Computer Science and Engineering, SASTRA Deemed University, Kumbakonam, Tamilnadu, India. (Email: meganathan@src.sastra.edu)
R.Bala Krishnan, Faculty of Computer Science and Engineering, SASTRA Deemed University, Kumbakonam, Tamilnadu, India.
A.Sumathi, Faculty of Computer Science and Engineering, SASTRA Deemed University, Kumbakonam, Tamilnadu, India.
S.Sheik Mohideen Shah, Faculty of Computer Science and Engineering, SASTRA Deemed University, Kumbakonam, Tamilnadu, India.
J.Senthilkumar Faculty of Computer Science and Engineering, SASTRA Deemed University, Kumbakonam, Tamilnadu, India.

Finally a total sum of 3545 data objects for NE monsoon months were presented for analysis. For a significant rainy day prediction a sum of 1404 rainy day instances were offered for examination. The procedures like discretization and Best-Fit (BF) ranges for

five synaptic features that are applied for the proposed mining process. The inputs for the above prescribed process are categorized as the following three bins namely, low_range, medium_range and high_range. The atmospheric class labeled "prcp" depicts the happening of the "rainfall" process, which is having the features "Yes" or "No" for wet / dry day. The presented Table 1 reports the nominal contents for the synaptic variables for the Chennai coastal region. The excerpted atmospheric data-set holds five attributes, their category and depiction which are stated in Table [2].

Table 1. Categorical data of Synaptic features for Chennai Location

Weather factor	Nominal value range	North East monsoon	
		24 hr advance	48 hr advance
Mean daily temperature in Fahrenheit	T_{LOW}	<76º(24.4ºC)	<76º(24.4ºC)
	T_{MED}	76º-83º	76º-83º
	T_{HIGH}	>83º(28.3ºC)	>83º(28.3ºC)
Dew point in Fahrenheit	D_{LOW}	<63º(17.2ºC)	<63º(17.2ºC)
	D_{MED}	63º-71º	63º-71º
	D_{HIGH}	>71º(21.6ºC)	>71º(21.6ºC)
Visibility in mile	V_{LOW}	<5.5(8.85kph)	<4.46(7.17kph)
	V_{MED}	5.5-9.7	4.46-7.63
	V_{HIGH}	>9.7(15.6kph)	>7.63(12.27kph)
Wind speed in knot	W_{LOW}	<7	<7
	W_{MED}	7-15	7-15
	W_{HIGH}	>15	>15
Precipitation	YES	>0	>0
	NO	≤0	≤0

Table 2. Dimension of weather data set.

Attribute	Type	Description
Year	Number	Year considered
Month	Number	Month considered
Day	Number	Day considered
TP	Number	Real mean temperature for the day in degrees Fahrenheit to tenths. (Celsius to tenths for metric version)
DP	Number	Real mean dew point for the day in degrees Fahrenheit to tenths. (Celsius to tenths for metric version).
WS	Number	Real mean wind speed for the day in knots to tenths. (Meters/second to tenths for metric version).
VISB	Number	Real mean visibility for the day in miles to tenths. (Kilometers to tenths for metric version).
RP	Number	Real total precipitation (rain) reported during the day in inches and hundredths. (millimeters to tenths for metric)

III. RESULT AND DISCUSSIONS

A) Wet and dry day weather patterns

The Association (ASS) Rule Mining has been productively established for forecast of rainfall for 24 hours prior to its occurrence (Sivaramakrishan and Meganathan 2011). The weather patterns are extracted by machine learning tool Weka 3 (www.cs.waikato.ac.nz) using the Predictive Apriori (Pre-Apriori) mechanism which generates the best association rules by giving its Support (Sup) and Confidence (Conf) values. Sup and Conf are two interestingness assesses of rule validity, which reflects the certainty and utility of identified patterns. Typically, ASS-Rules are concerning if they satisfy both a Min- Sup threshold and a Min- Conf threshold. The weather patterns are associated with the atmospherically parametric daily Mean Temp, DP, WS and VISB. The class label attribute of the ASS-Rule rule for wet or dry day prediction is "PRCP" (ie. precipitation). If class label "PRCP" is "yes" then the association rule is considered for occurrence of rainy day or wet day. If the value of class label "PRCP" is "no" then the ASS rule is considered for non occurrence of rainy day or dry day. Best Ass-Rules are considered for the prediction by applying the given dataset for identifying the occurrence and nonoccurrence of rain on dry and wet days for 24 hours ahead [Table 3] and 48 hours ahead [Table 4] during Northeast monsoon periods of Chennai station.

Table 3. Patterns of Weather for 24 hour betterment of RF prediction at some point in NE monsoon.

ASS Rule ($X \Rightarrow Y$)	SUP($X \cup Y$)	CONF P(Y/X)
DEWP='(-inf-63.1]' VISIB='(-inf-5.533333]' WDSP='(-inf-7.333333]' ==> PRCP=no	69	0.99454
TEMP='(75.966667-83.033333]' DEWP='(-inf-63.1]' ==> PRCP=no	40	0.99352
TEMP='(-inf-75.966667]' DEWP='(63.1-71.3]' VISIB='(5.533333-9.766667]' WDSP='(-inf-7.333333]' ==> PRCP=no	12	0.97576
TEMP='(75.966667-83.033333]' DEWP='(63.1-71.3]' VISIB='(5.533333-9.766667]' ==> PRCP=no	19	0.91845
TEMP='(-inf-75.966667]' VISIB='(9.766667-inf)' ==> PRCP=no	4	0.91683
DEWP='(71.3-inf)' WDSP='(14.666667-inf)' ==> PRCP=yes	4	0.91683
TEMP='(-inf-75.966667]' DEWP='(71.3-inf)' WDSP='(-inf-7.333333]' ==> PRCP=yes	129	0.8814
TEMP='(75.966667-83.033333]'WDSP='(14.666667-inf)'==> PRCP=yes	2	0.86478
TEMP='(-inf-75.966667]' VISIB='(-inf-5.533333]' WDSP='(7.333333-14.666667]' ==> PRCP=yes	36	0.77525
TEMP='(75.966667-83.033333]' DEWP='(71.3-inf)' WDSP='(7.333333-14.666667]' ==> PRCP=yes	37	0.70204

IV. VALIDATION

Validation for the data mining principles has been enforced using the CV(Cross-Validation) and STS (Supplied-Test-Set) mechanisms.

Retrieval Number C6802098319/2019©BEIESP
DOI: 10.35940/ijrte.C6802.098319
Journal Website: www.ijrte.org

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication

The skill of a Machine Learning (ML) procedure proposes that by employing ten equal quantity partitions habitually afford the identical error_rate (ER) as if the entire data had been applied for training phase.

Table 4. Association Rules for 48-hour prior prediction during NE monsoon

ASS Rule ($X \Rightarrow Y$)	SUP ($X \cup Y$)	CONF P(Y/X)
TEMP = '(75.966667-83.033333]' DEWP='(-inf-63.1]' VISIB='(4.466667-7.633333]' WDSP='(-inf-7.333333]' ==> PRCP=no	18	0.98344
TEMP='(-inf-75.966667]' DEWP='(63.1-71.3]' VISIB='(4.466667-7.633333]' ==> PRCP=no	49	0.91209
TEMP='(75.966667-83.033333]' DEWP='(-inf-63.1]' WDSP='(-inf-7.333333]' ==> PRCP=no	38	0.88434
TEMP='(-inf-75.966667]' DEWP='(63.1-71.3]' WDSP='(-inf-7.333333]' ==> PRCP=no	296	0.86668
TEMP='(-inf-75.966667]' DEWP='(71.3-inf)' VISIB='(-inf-4.466667]' WDSP='(-inf-7.333333]' ==> PRCP=yes	125	0.76302
TEMP='(-inf-75.966667]' DEWP='(71.3-inf) ==>PRCP=yes	146	0.72907
TEMP='(83.033333-inf)' VISIB='(-inf-4.466667]' WDSP='(7.333333-14.666667]' ==> PRCP=no	5	0.69593
TEMP='(83.033333-inf)' VISIB='(4.466667-7.633333]' WDSP='(-inf-7.333333]' ==> PRCP=no	148	0.68543
TEMP='(75.966667-83.033333]' DEWP='(71.3-inf)' VISIB='(-inf-4.466667]' WDSP='(7.333333-14.666667]' ==> PRCP=yes	28	0.66696

In STS scheme, dataset of 45 years from 1961-2005 is considered as training-set data and residual individual years 2006, 2007, 2008, 2009 and 2010 are used as testing-set respectively. The validation process is accomplished to detect the reliability of the rendered outcomes and to confirm whether they can be applied in real time environment with sensitive input data for the prediction of dry and wet days over the required period of time. The stratified Ten-fold CV outcomes are stated for the prediction of wet / dry day during NE monsoon periods in [Table 6]. The confusion matrix is also obtained for the above forecasting is shown in [Table 7]. For the predicting occurrence and nonoccurrence of the wet and dry days, the ML principle K* achieves acceptable success rates using CV method for 24 hours ahead and 48 hours ahead respectively The validation results are presented in [Table 5] using STS scheme for wet / dry day forecasting and considerable rainy day forecasting. The exactness of classification system is anticipated through the Ten fold CV mechanism

Table 5. Summary of success rate using supplied test set method.

Assessment	No. of samples	Prediction type	Correlation coefficient	Testing years				
				2006	2007	2008	2009	2010
			%	%	%	%	%	%
NE monsoon	3544	24hr prior	69.89	79.55	66..66	66.66	77.27	67.04
	3545	48hr prior	65.13	74.71	59.30	73.26	72.41	68.97

Table 6. Classification accuracy using 10-fold cross validation for rainfall prediction.

Cross-Validation (Stratified)	North East monsoon months			
	Prior to: 24 Hrs		Prior to: 48 Hrs	
	Instances	%	Instances	%
Properly Classified Instances	2477	69.9	2309	65.1
Incorrectly Classified Instances	1067	30.1	1236	34.9

Table 7. Confusion matrix of 10-fold cross validation for rainfall prediction.

Monsoon	Northeast monsoon			
Prediction type	Prior to: 24 Hrs		Prior to: 48 Hrs	
classified as →	No	Yes	No	Yes
Actual class "No"	1436	706	1453	688
Actual class "Yes"	361	1041	548	856
Success rate in %	69.9 %		65.1 %	

A) Statistical Summary of South West Monsoon Precipitation Predictor Model for Chennai Station

Evaluation Results on user defined year wise test set for rainfall estimation with the threshold values of 20 millimeter for Basic Estimation Model is presented in Table 8. The statistical measures correctly classified instances (CCI),incorrectly classified instances (ICI), kappa statistic (KP), mean absolute error (MAE), root mean squared error (RMSE), relative absolute error (RAE), root relative squared error (RRSE), total number of instances (TNI)are evaluated on the testing data set on the following years 2006 to 2010. The total number of instances of the training data set from the year of 1951 – 2005 is 1402 which is taken from SE monsoon months of June, July, August and September of the concerned station after the preprocessing techniques.

Table 8. Basic estimation model for rainfall prediction.

Statistical Measures	**2006**	**2007**	**2008**	**2009**	**2010**
CCI	73 %	83	70	63	85
ICI	27 %	17	30	37	15
KP	0.000	0.000	0	0	0.191
MAE	0.375	0.2924	0.3572	0.4154	0.3643
RMSE	0.4304	0.3508	0.4364	0.4632	0.3923
RAE	96.7379%	87.2312%	89.3564%	95.5022	107.5493
RRSE	96.553 %	91.7231%	95.0259%	93.3826	101.3334
TNI	44	36	37	38	46

Classifier accuracy measures for the rainfall estimation with the threshold value of 20 millimeter are shown in the following Table 8. The classifier accuracy measures are obtained in the prescribed year data set based on the user defined testing data for the two different classes. First one is "low" means the rainfall is low if the precipitation occurs below the 20 millimeter. The second class label "high" describes the rainfall is high if the precipitation occurs greater than 20 millimeter. The true positive rate (TPR), false positive rate (FPR), precision (PR), recall (RC), F-measure (FM), ROC area (ROCA) and weighted average (WA) classifier measures are obtained for the two class labels "low" and "high" for the rainfall estimation during the SW monsoon months over the station.

The evaluation results for basic prediction on user defined year wise test set for the occurrence of the rainfall with the threshold values of 0 millimeter is shown in the following Table 9.

Table 8. Rainfall estimation with the threshold value of 20 millimeter

YEAR	2006			2007			2008			2009			2010		
CLASS	LOW	HIGH	WA	LOW	HIGH	WA	LOW	HIGH	WA	LOW	HIGH	WA	LOW	HIGH	WA
TPR	1	0	0.727	1	0	0.833	1	0	0.703	1	0	0.632	1	0.125	0.848
FPR	1	0	0.727	1	0	0.833	1	0	0.703	1	0	0.632	0.875	0	0.723
PR	0.727	0	0.529	0.833	0	0.694	0.703	0	0.494	0.632	0	0.399	0.844	1	0.871

RC	1	0	0.727	1	0	0.833	1	0	0.703	1	0	0.632	1	0.125	0.848
FM	0.842	0	0.612	0.909	0	0.758	0.825	0	0.58	0.774	0	0.489	0.916	0.222	0.795
ROCA	0.654	0.654	0.654	0.761	0.761	0.761	0.764	0.764	0.784	0.667	0.667	0.667	0.651	0.651	0.651

Table 9. Evolution of Rainfall occurrence

Statistical Measures	10-Fold Cross Validation	2006	2007	2008	2009	2010
CCI	64	59	59	63	61	55
ICI	36	41	41	37	39	45
KP	0.0935	0.0476	0.257	0.0288	0.0596	0.132
MAE	0.4506	0.4807	0.4672	0.4857	0.4603	0.4762
RMSE	0.4722	0.4912	0.4753	0.4899	0.4749	0.4853
RAE	95.5832	100.1547	90.1868	103.9631	96.7626	94.0777
RRSE	97.2664	98.6846	88.8863	101.0539	96.233	92.721
TNI	4681	113	113	114	110	109

The consolidated results for the correlation coefficient results are shown in the Table 9 for the various prediction types for the occurrence of the precipitation over the mentioned station such as 24 hour prediction (Basic Prediction), 48 hour prediction (Advanced Prediction), 24 hour prediction with estimation of low precipitation and high precipitation with threshold value of 20 millimeter (Basic Estimation) and 48 hour prediction (Advanced Estimation) with estimation of low and high precipitation with the same 20 millimeter threshold value for the class label values “LOW” and “HIGH”. The class label values for the basic prediction and advanced prediction are “YES” and “NO” for the occurrence of the precipitation and non occurrence of the precipitation respectively.

The performance rate (55%)of precipitation prediction over Chennai station for the typical southwest monsoon months in the testing years 2006 to 2010 by the proposed data mining model (PDMM) is compared with the performance rate (54%)of the existing models namely India Meteorological Department observed synoptic model (IMDSM) of National Centre for Medium Range Weather Forecasting, Pune, India (NCMRWF) and statistical model performance rate (46%) based on past 100 years moving average data model (SM) which is generated by ARIMA model , that is Auto Regressive Integrated Moving Average process. It is seen that while the PDMM method gives a success percentage over the others.

Table 9. Success rate of SW monsoon prediction rates.

S.No.	Prediction Type	No. of samples	Co-relation Co-efficient	Testing years				
				2006	2007	2008	2009	2010
			%	%	%	%	%	%
1	Basic Prediction	4681	64.49	59	59	63	60	55
2	Advanced Prediction	4682	62.66	58	52	62	62	53
3	Basic Estimation	4681	74.59	73	83	70	63	85
4	Advanced Estimation	4682	62.64	71	75	68	65	79

V. CONCLUSION

The Ass-Rule mining of DM technique and object based classifier principle has been subjected to predict the incidence of wet and dry days with class labels. The observed results states that, the forecasting of Ass-Rule mining is practically precise and the model is suitable for predicting the occurrence and non occurrence of the rainy day during NE and SW monsoon months for Chennai station. As grounds in the observations, the presented models is appropriate for supervising the weather conditions and detects the rainy days 48 hours ahead and estimate the normal and high precipitation. The proposed method assures to be a suitable one for tropical coastal regions.

VI. ACKNOWLEDGEMENTS

Thanks to SASTRA Deemed University for providing Discrete Mathematics Laboratory funded by Department of Science and Technology located at Srinivasa Ramanujan Centre, Kumbakonam, Tamil Nadu, India, to develop the proposed model and obtain the results.

REFERENCES

1. Agrawal, R. and Srikant, R., 1994, "Fast Algorithms for Mining Association Rules", Proc. of the 20th Int. Conference on Very Large Databases, Santiago, Chile, Sept. 1994. Expanded version available as IBM Research Report RJ9839.
2. Agrawal, R. Mannila, H. Srikant, R. Toivonen, H. and Verkamo, A.I.,1995, "Fast Discovery of Association Rules", Advances in Knowledge Discovery and Data Mining, Chapter 12, AAAI/MIT Press.
3. Cleary, J.G. and Trigg, L.E., 1995, "K*: An Instance-based learner using an entropic distance measure", Proceedings of the 12th Int. Conf. on Machine Learning, San Francisco, Morgan Kaufmann, 108-114.
4. Balachandran, S., Asokan, R. and Sridaran, S., 2006, "Global surface temperature in relation to northeast monsoon rainfall over Tamil Nadu", Journal of Earth System Science, 115, 3, 349-362.
5. Mohanty, V.C., 1994, "Forecast of precipitation over Delhi during SW Monsoon", Mausam, 45, 87p.
6. Rao Krishna, P.R. and Jagannathan, P., 1953, "A study of the northeast monsoon rainfall of Tamilnadu", Indian Journal of Meteorology and Geophysics, 4, 22-43.
7. Sivaramakrishnan, T.R., 1988, "Rainfall characteristics of Sriharkota", ISRO Scientific Report No. 05-026-88.
8. Sivaramakrishnan, T.R., 1989, "Annual rainfall over Tamil Nadu", Hydrology Journal, IAH, 20p.
9. Sivaramakrishnan, T.R. and Meganathan, S., 2011, "Association Rule Mining and Classifier Approach for Quantitative Spot Rainfall Prediction", Journal of Theoretical and Applied Information Technology, 34(2), 173-177.
10. Sivaramakrishnan, T.R. and Meganathan, S., 2012, "Data mining as a tool for precipitation prediction, Archives Des Sciences, 65(3), 8.
11. Sivaramakrishnan, T.R. and Meganathan, S., 2012, "Point rainfall prediction using data mining technique", Res. Journal of App. Sciences, Engineering and Technology, 4(13), 1899-1902.
12. Sivaramakrishnan, T.R. and Meganathan, S., 2013, "Association rule mining and classifier approach for 48 – hour rainfall prediction over Cuddalore station of east coast of India", Res. Journal of App. Sciences, Engineering and Technology, 5(14), 3692-3696.
13. Sivaramakrishnan, T.R. and Meganathan, S., 2014, "A Technique for Spot Forecasting", Mausam. Vol. 65, Article No. J-065(5555). (Accepted for publication).
14. Zubair, L. and Ropelewski, 2006, "The strengthening relationship of ENSO and the North East Monsoon rainfall over Sri Lanka and Southern India", Journal of Climate, 19(8), 1567-1575.