MAUSAM, **66**, 1 (January 2015), 33-42

551.509.3

A technique for spot weather forecasting

S. MEGANATHAN and T. R. SIVARAMAKRISHNAN

Srinivasa Ramanujan Centre, SASTRA University, Kumbakonam, Tamilnadu – 612 001, India

(*Received 20 June 2013, Modified 11 September 2013*)

e mail : vsmnathan@gmail.com

सार – इस शोध पत्र का उद्देश्य दक्षिण भारत के अंतरदेशीय केंद्र तिरूचिरापल्ली (10° 46′ उ. अक्षांस / 78° 43′ पू. देशांतर) में उत्तर-पूर्व और दक्षिण-पश्चिमी मॉनसून ऋतु के दौरान 24 घंटे और 48 घंटे पूर्व होने वाली संभावित वर्षा (वर्षा वाले दिन) के बारे में प्रचालनात्मक पूर्वानुमानकर्ताओं का मार्गदर्शन करना है। इस पद्धति की दिलचस्प बात यह है कि यह मौसम विज्ञानिक प्राचलों के केवल उसी स्थान विशेष के आँकड़ों पर आधारित है। इसमें आँकड़ा माइनिंग तकनीक का उपयोग किया गया है। इसी पद्धति का उपयोग करते हुए ग्रीष्म ऋतु में अप्रैल और मई के महीनों तथा शीत ऋतु में जनवरी, फरवरी और मार्च के महीनों के क्रमश: गर्म दिन और ठंडे दिन के मौसम के पैटर्नों का भी पता लगाया गया है। इस अध्ययन से पता चला कि प्रस्तावित आँकड़ा माइनिंग मॉडल, स्थानीय मौसम प्राचलों की मदद से ग्रीष्म ऋतु के दौरान वर्षा, गर्म दिन तथा शीत ऋतु के दौरान ठंडे दिन का पूर्वानुमान कर सकता है।

ABSTRACT. The emphasis of present work is to provide some guidance to the operational forecasters for indicating the possible occurrence of rainfall (wet day) before 24 hours and 48 hours during Northeast and Southwest monsoon season over the inland station Trichirappalli (Latitude 10°46' N / Longitude 78°43' E) of south India. The interesting aspect of the methodology is that it is based only on the "*in situ*" data of meteorological parameters. Data mining technique is used. The weather patterns of hot day and cold day for April and May months during summer and January, February and March months during winter respectively have also been extracted using the same method. The study shows that the proposed data mining model can predict the occurrence of rainfall, hot day during summer and cold day during winter with the help of local weather parameters.

Key words – Association rule, Data mining, Weather pattern, Weather forecasting.

1. Introduction

Weather forecast is mostly for a region based on synoptic situations. The forecast model may be either synoptic or numerical one (Balachandran *et al*., 2006; Dhar and Rakhecha, 1983; Kripalani and Kumar, 2004; Kumar *et al*., 2007; Mohanty *et al*., 2001; Nayagam Lorna, 2009; Raj, 1996; Rao, 1963; Shukla and Mooley, 1987; Sivaramakrishnan and Sreedharan 1988,1989; Sridaran and Muthuchami, 1990; Zubair and Ropelewski, 2006). The models need the input of meteorological parameters at different grid points. During the passage of clean tropical systems, sequence of events is known and mostly the knowledge of the sequence is helpful for predicting the weather. Cyclones and summer monsoon activity (Mukherjee and Sivaramakrishnan, 1977; Rao, 1976; Raghavan *et al*., 1983; Sivaramakrishnan and Prasad, 1999) are the few systems with clear cut rainfall patterns over any place. A few investigations on the nature and trend of point rainfall as well as its prediction (Seetharam, 2009; Sivaramakrishnan, 1988) are also seen. However need for forecast at specific location is increasing day by day. Industrial townships, space stations, harbours etc. need this. Again metropolitan cities also seek such spot forecast for national festivals, sports and similar occasions.

Operationally point rainfall (RF) forecast is given to a very limited extent. Nevertheless radar and satellite data of the recent decades help in assessing the rain potential of clouds (Raghavan *et al*., 1987). However this type of rain forecast is effective for 12 hours or less. A useful forecast has to be at least 24 hours ahead. If it is 48 hours ahead that is preferable for operational exercise. Recently the authors have successfully developed some methods to forecast wet and dry days based on data mining techniques for some coastal stations of east coast of India (Meganathan *et al*., 2009; Sivaramakrishnan and Meganathan 2011, 2012a, 2012b, 2012c, 2013). In the present paper, a similar model is tried for a sample inland station, Trichirappalli of south India.

TABLE 1

Nominal values for atmospheric parameters

Weather parameter	Nominal variable	Nominal values			
		24 hr NE	48 hr NE	24 hr SW	48 hr SW
Temperature (Fahrenheit)	T_L	< 75.13	< 75.56	< 81.66	< 81.93
	T_M	75.13 - 82.46	75.56 - 83.16	81.66 - 88.83	81.93 - 89.36
	T_H	> 82.46	> 83.16	> 88.83	> 89.36
Dew Point (Fahrenheit)	D_L	< 60	< 60	< 63.03	< 63.03
	D_M	60-68.9	60 - 68.9	63.03 - 70.86	63.03 - 70.86
	D_H	> 68.9	> 68.9	> 70.86	> 70.86
Wind Speed (Knots)	W_L	< 8.53	< 8.53	< 10.53	< 10.53
	W_M	8.53 - 17.06	8.53 - 17.06	10.53 - 20.76	10.53 - 20.76
	W_H	> 17.06	> 17.06	> 20.76	> 20.76
Visibility (Miles)	V_L	< 5.13	< 5.13	< 5.2	< 5.2
	V_M	5.13 - 9.86	5.13 - 9.86	5.2 - 8.6	5.2 - 8.6
	V_H	> 9.86	> 9.86	> 8.6	> 8.6
Precipitation (Inches)	YES	> 0	> 0	> 0	> 0
	NO	= 0	= 0	= 0	= 0

2. Data used

Trichirappalli (Latitude 10°46' N / Longitude 78° 43' E) is located in Tamil Nadu state of South India which is a metropolitan area with an airport. India Meteorological Department (IMD) has a full fledged observatory here. Rainy days during southwest (SW) monsoon and northeast (NE) monsoon as well as minimum temperature during winter and maximum temperature during summer have been taken as the parameters for the forecast because of their operational importance. Data used are those recorded by IMD and then included in World Meteorological Organization (WMO) data bank as managed by the National Oceanic and Atmospheric Administration (NOAA), which is a federal agency focused on the condition of the oceans and the atmosphere worldwide. Daily rainfall data for the years 1961-2010 were considered for developing the model.

Our data set consists of five atmospheric variables including temperature, dew point, wind speed, visibility and precipitation (rainfall). The data set that we extracted consists of the prevailing atmospheric situations 24 hours and 48 hours before the actual occurrence of the rain during the North East monsoon months of October, November and December as well as the South West monsoon months of June, July, August and September. Data preprocessing steps were applied on the raw set of seasonal data and they were converted to nominal values by applying filtering using unsupervised attribute of discretization algorithm. After the filtering operations were carried out, a total of 3393 instances for 24 hours advance wet day prediction in NE monsoon, a total of 3398 instances for 48 hours advance wet day prediction and 4609 and 4614 instances respectively for 24 hours advance prediction and 48 hours advance prediction during SW monsoon were present for analysis. The discretization algorithm produced various best-fit ranges for the five atmospheric conditions we used in analysis. Based on the discretization algorithm the atmospheric variables are ranged into the nominal values low, medium and high with its best ranges using machine learning tool Weka 3.6. The nominal values for the atmospheric parameters are shown in Table 1.

3. Methodology

3.1. *Association rule mining for prediction*

The problem of mining association rule was first introduced in the last decade by database communities (Agrawal and Srikant, 1994; Agrawal *et al.*, 1995). Association rule mining aims to extract the interesting correlations, associations, frequent patterns among set of weather items in climate data repository. For the spot specific forecast, the derived frequent weather patterns are

TABLE 2

Generated association rules for 24 hours ahead rainfall prediction during SW monsoon with support and confidence values

Association Rule $(A \Rightarrow B)$	Support $(A \cup B)$	Confidence P(B/A)
TEMP = '(-inf-81.666667)' DEWP = '(70.866667-inf)' VISIB = '(5.2-8.6]' WDSP = '(-inf-10.533333]' $\Rightarrow$ PRCP = yes	18	0.77891
TEMP = '(-inf-81.666667) DEWP = '(63.033333-70.866667]' VISIB = '(-inf-5.2]' $\Rightarrow$ PRCP = no	20	0.75277
TEMP = '(88.833333-inf)' DEWP = '(63.033333-70.866667]' VISIB = '(5.2-8.6]' WDSP = '(10.533333-20.766667]' $\Rightarrow$ PRCP = no	161	0.93478
TEMP = '(81.666667-88.833333]' WDSP = '(20.766667-inf)' $\Rightarrow$ PRCP = no	124	0.98943
TEMP = '(88.833333-inf)' DEWP = '(63.033333-70.866667]' WDSP = '(-inf-10.533333]' $\Rightarrow$ PRCP = no	108	0.9087
TEMP = '(88.833333-inf)' DEWP = '(70.866667-inf)' WDSP = '(10.533333-20.766667]' $\Rightarrow$ PRCP = no	164	0.87631
DEWP = '(70.866667-inf)' VISIB = '(8.6-inf)' WDSP = '(-inf-10.533333]' $\Rightarrow$ PRCP = yes	8	0.66046
TEMP = '(-inf-81.666667]' VISIB = '(5.2-8.6]' WDSP = '(-inf-10.533333]' $\Rightarrow$ PRCP = yes	25	0.65351

represented in the form of association rules. These information leads to the discovery of interesting associations and correlations within the data with the help of support and confidence measures for the occurrence of the weather event. Recently the authors (Sivaramakrishnan and Meganathan, 2013) have reported the suitability of association rule approach for point rainfall prediction 24 hours ahead in a case study. When we apply the above association rule concept to studying meteorological data, with each record listing various atmospheric observations including wind direction, wind speed, temperature, relative humidity, rainfall and mean sea level pressure taken at a certain time in certain area we can find association rules like,

R_1: If the humidity is medium wet, then there is no rain in the same area at the same time.

Although rule R_1 reflects some relationships among the meteorological elements, its role in weather prediction is inadequate. The association rule has been used in climatology applications with more number of parameters. The following rule emphasis the occurrence of the weather event with available atmospheric conditions.

R_2: If the wind direction is east and the weather is warm, then it keeps warm for the next 48 hour.

R_3: If temperature is low, dew point is high, wind speed is medium then precipitation has occurred.

For example, a derived weather pattern for the wet day prediction, Temp is “T_{LOW}” and Dew point is “D_{HIGH}” and Visibility is “V_{MEDIUM}” and Wind speed is “W_{LOW}” that emphasis the occurrence of the wet day. The temperature, dew point, visibility and wind speed are the input atmosphere parameters. If the values for these variables would be within their nominal value range as shown in Table 1, which is fixed by the machine learning tool, then precipitation will be occurred. The class label “precipitation” describes it. We used predictive Apriori algorithm for deriving the association rule from the preprocessed dataset. The basic property of Apriori is that all non-empty subsets of a frequent item set must be frequent. In connection with the above the predictive Apriori algorithm searches with an increasing support threshold for the best 'n' rules concerning a support-based corrected confidence value.

Predictive mining is a task that it performs inference on the current data in order to make a prediction. Here the climate parameters temperature, dew point, visibility, wind speed and precipitation are taken for analysis using classification and association mining. The rule $A \Rightarrow B$ holds in the transaction set D with support s, where s is the percentage of transactions in D that contain $A \cup B$ (*i.e.*, the union of sets A and B, or say, both A and B). This is taken to be the probability, P ($A \cup B$). The rule $A \Rightarrow B$ has confidence c in the transaction set D, where, c is the percentage of transactions in D containing A that

TABLE 3

Generated association rules for 48 hours ahead rainfall prediction during SW monsoon with support and confidence values

Association Rule $(A \Rightarrow B)$	Support $(A \cup B)$	Confidence P(B/A)
TEMP = '(89.366667-inf)' DEWP = '(63.033333-70.866667]' VISIB = '(-inf-5.2]' WDSP = '(10.533333-20.766667]' ⇒ PRCP = no	261	0.88738
TEMP = '(-inf-81.933333]' DEWP = '(63.033333-70.866667]' WDSP = '(10.533333-20.766667]' ⇒ PRCP = no	12	0.86619
TEMP = '(89.366667-inf)' DEWP = '(70.866667-inf)' WDSP = '(10.533333-20.766667]' ⇒ PRCP = no	122	0.83411
TEMP = '(81.933333-89.366667]' DEWP = '(63.033333-70.866667]' WDSP = '(10.533333-20.766667]' ⇒ PRCP = no	1402	0.65042
TEMP = '(-inf-81.933333]' DEWP = '(63.033333-70.866667]' VISIB = '(5.2-8.6]' ⇒ PRCP = yes	11	0.4998
TEMP = '(-inf-81.933333]' DEWP = '(70.866667-inf)' VISIB = '(5.2-8.6]' WDSP = '(-inf-10.533333]' ⇒ PRCP = yes	20	0.4762
TEMP = '(-inf-81.933333]' DEWP = '(70.866667-inf)' VISIB = '(-inf-5.2]' WDSP = '(-inf-10.533333]' ⇒ PRCP = yes	185	0.46316

TABLE 4

Generated association rules for 24 hours ahead rainfall prediction during NE monsoon with support and confidence values

Association Rule $(A \Rightarrow B)$	Support $(A \cup B)$	Confidence P(B/A)
TEMP = '(75.133333-82.466667]' DEWP = '(60-68.9]' VISIB = '(5.133333-9.866667]' WIND = '(-inf-8.533333]' ⇒ PRCP = yes	31	0.99155
TEMP = '(82.466667-inf)' DEWP = '(60-68.9]' VISIB = '(-inf-5.133333]' WIND = '(8.533333-17.066667]' ⇒ PRCP = yes	24	0.94504
TEMP = '(-inf-75.133333]' DEWP = '(68.9-inf)' WIND = '(8.533333-17.066667]' ⇒ PRCP = no	34	0.8335
TEMP = '(-inf-75.133333]' WIND = '(17.066667-inf)' ⇒ PRCP = no	3	0.89357
TEMP = '(-inf-75.133333]' DEWP = '(68.9-inf)' WIND = '(8.533333-17.066667]' ⇒ PRCP = no	34	0.8335

also contain B. This is taken to be the conditional probability, P (B|A). That is,

$$\text{Support } (A \Rightarrow B) = P(A \cup B)$$

$$\text{Confidence } (A \Rightarrow B) = \frac{\text{Support\_count } (A \cup B)}{\text{Support\_count } (A)}$$

Recently association rule mining was successfully verified for prediction of rainfall just 24 hours before (Sivaramakrishnan and Meganathan, 2011). The predictive Apriori algorithm shows the best association rules. Some of the best rules that have been predicted from the given dataset are shown in Table 2 for 24 hour advance rainfall prediction during SW monsoon, Table 3 for 48 hour advance rainfall prediction during SW monsoon and Table 4 for 24 hour advance rainfall prediction during NE monsoon. Each and every association rule consists with a support and confidence value that determines the credibility of the rule.

3.2. *Classification*

Classification is a learning function that maps the data into one of the several predefined data classes. It is a

TABLE 5

Statistical summary of rainfall prediction on SW and NE monsoon

Measures	Southwest monsoon		Northeast monsoon	
	24 hour before	48 hour before	24 hour before	48 hour before
Total number of instances	4609	4614	3393	3394
Correctly classified instances	3712	3698	2254	2189
Incorrectly classified instances	897	916	1139	1205
Correctly classified in %	80.5381	80.1474	66.43	64.5
Incorrectly classified in %	19.4619	19.8526	33.57	35.5
Mean absolute error	0.2759	0.2948	0.4197	0.4673
Root mean squared error	0.367	0.3818	0.4503	0.4807

TABLE 6

A confusion matrix for positive and negative tuples of precipitation prediction

Actual class	Predicted class	
	Precipitation = YES	Precipitation = NO
Precipitation = YES	True positives	False negatives
Precipitation = NO	False positives	True negatives

TABLE 7

Confusion matrix for wet/dry day prediction on SW and NE monsoon

Monsoon	Prediction type	Confusion matrix values					
		True positives	False negatives	False positives	True negatives	No. of instances	Correlation coefficient
Southwest	24 hr advance	81	833	64	3631	4609	80.5 %
	48 hr advance	0	916	0	3698	4614	80.1 %
Northeast	24 hr advance	2175	18	1121	79	3393	66.4 %
	48 hr advance					3394	64.5 %

form of data analysis that can be used to extract models describing important class to predict future data trends. It predicts on categorical labels like occurrence of wet day, occurrence of cool day, occurrence of hot day etc. Classification is used to predict the data classes whose class label is unknown. Here we use K* classification algorithm (Cleary and Trigg, 1995) which is an instance based classifier, that is the class of a test instance is based upon the class of those training instances similar to it, as determined by some similarity function, such as entropy based similarity function. By using this, the discretized data of the atmospheric situations before 24 hours and 48 hours of the actual rainy day was evaluated and the coherence of correctly classified instances and incorrectly classified instances were found out to justify the accuracy of the data prediction model we used.

In this paper the data class "PRCP" is used for wet and dry day forecasting during SW and NE monsoon season of sample station that has two values "yes" and "no". The value "yes" in the derived pattern of the association rule describes the occurrence of the wet day. The value "no" in the derived pattern of the association rule describes the occurrence of the dry day. Similarly the forecasting of the cool day during winter months, the derived pattern of the association rule has the class label "POST_MIN" which has two values "nor" and "low". The class label value "nor" is meant for the normal winter day

TABLE 8

Detailed accuracy by Class for precipitation prediction on SW and NE monsoon

Prediction type	No. of samples	Correlation coefficient	Class	TP rate	FP rate	Precision	Recall	F-measure	ROC area
24 hour advance in NE monsoon	3393	66.4309	NO	0.666	0.008	0.814	0.066	0.122	0.689
			YES	0.992	0.934	0.66	0.099	0.792	0.689
			Wg. Avg	0.664	0.607	0.715	0.664	0.555	0.689
48 hour advance in NE monsoon	3394	64.4909	NO	0.617	0.316	0.717	0.617	0.663	0.669
			YES	0.684	0.383	0.579	0.684	0.627	0.669
			Wg. Avg	0.646	0.345	0.657	0.646	0.648	0.669
24 hour advance in SW monsoon	4609	80.5381	NO	0.983	0.911	0.813	0.983	0.89	0.758
			YES	0.089	0.017	0.559	0.089	0.153	0.758
			Wg. Avg	0.805	0.734	0.763	0.805	0.744	0.758
48 hour advance in SW monsoon	4609	80.1474	NO	1	1	0.801	1	0.89	0.695
			YES	0	0	0	0	0	0.695
			Wg. Avg	0.801	0.801	0.642	0.801	0.713	0.695

TABLE 9

Success rate of wet/dry day prediction on SW and NE monsoon

Monsoon	Prediction type	No. of samples	Correlation coefficient (%)	Testing years				
				2006	2007	2008	2009	2010
				(%)	(%)	(%)	(%)	(%)
Southwest	24 hr prior	4609	80.5381	76	68	80	85	66
	48 hr prior	4609	79.5411	77	68	80	85	64
Northeast	24 hr prior	3393	66.4	56	69	73	63	47
	48 hr prior	3394	64.5	56	69	63	43	59

and “low” is for the extreme winter day (*i.e.*, cool day). The class label “POST_MAX” is for the hot day prediction during summer months whose values are “nor” and “high” which describe normal summer day and extreme summer day (*i.e.*, hot day) respectively.

4. Results and discussion

4.1. *Validation*

Validation for our model has been done using the 10-fold cross validation, percentage split method and supplied test method. The basic notions of those methods have been described here. Classifiers rely on being trained before they can reliably be used on new data. Of course, it stands to reason that the more instances the classifier is exposed to during the training phase, the more reliable it will be as it has more experience. However, once trained, we would like to test the classifier too, so that we are confident that it works successfully. For this, yet more unseen instances are required.

A problem that often occurs is the lack of readily available training/test data. These instances must be pre-classified which is typically time-consuming. A method to circumvent this issue is known as cross-validation. It works as follows:

(*i*) Separate data in to fixed number of partitions (or folds).

(*ii*) Select the first fold for testing, whilst the remaining folds are used for training.

TABLE 10

Generated association rules for 24 hours before cold day prediction during winter months with support and confidence values

Association Rule $(A \Rightarrow B)$	Support $(A \cup B)$	Confidence P(B/A)

MAX = '(82.8-91.8]' MIN = '(60.2-70.4]' DEWP = '(-inf-58.866667]'
VISIB = '(-inf-4.7]' ⇒ POST_MIN = low
	15	0.98572

MAX = '(91.8-inf)' WDSP = '(9.533333-18.266667]' ⇒ POST_MIN = nor	14	0.98487
MIN = '(70.4-inf)' DEWP = '(67.733333-inf)' WDSP = '(9.533333-18.266667]' ⇒ POST_MIN = nor 131	134	0.97931
MAX = '(-inf-82.8]' DEWP = '(67.733333-inf)' 77 ⇒ POST_MIN = nor 75	77	0.97789
MAX = '(82.8-91.8]' DEWP = '(-inf-58.866667]' VISIB = '(-inf-4.7]' 17 ⇒ POST_MIN = low 16	17	0.94013

MAX = '(82.8-91.8]' MIN = '(60.2-70.4]' DEWP = '(-inf-58.866667]'
WDSP = '(-inf-9.533333]' 15 ⇒ POST_MIN = low 14
	15	0.92932

TABLE 11

Generated association rules for 48 hours before cold day prediction during winter months with support and confidence values

Association Rule $(A \Rightarrow B)$	Support $(A \cup B)$	Confidence P(B/A)
MAX = '(91.8-inf)' MIN = '(70.4-inf)' WDSP = '(-inf-9.533333]' ⇒ POST_MIN = nor	206	0.99497
MAX = '(91.8-inf)' DEWP = '(67.733333-inf)' WDSP = '(-inf-9.533333]' ⇒ POST_MIN = nor	174	0.99494

MAX = '(82.8-91.8]' MIN = '(70.4-inf)' VISIB = '(-inf-5.566667]'
WDSP = '(9.533333-18.266667]' ⇒ POST_MIN = nor
	150	0.95884

MAX = '(82.8-91.8]' MIN = '(60.2-70.4]' DEWP = '(-inf-58.866667]'
WDSP = '(-inf-9.533333]' ⇒ POST_MIN = low
	15	0.78154

MAX = '(82.8-91.8]' DEWP = '(-inf-58.866667]' WDSP = '(9.533333-18.266667]' ⇒ POST_MIN = low	4	0.74481

TABLE 12

Generated association rules for 24 hours before hot day prediction during summer months with support and confidence values

Association Rule $(A \Rightarrow B)$	Support $(A \cup B)$	Confidence P(B/A)
MAX = '(90.8-104.6]' DEWP = '(72.533333-inf)' VISIB = '(6.766667-11.333333]' ⇒ POST_MAX = nor	22	0.98956
MIN = '(71.733333-80.666667]' DEWP = '(72.533333-inf)' WDSP = '(12.2-23.5]' ⇒ POST_MAX = nor	20	0.98851

MAX = '(90.8-104.6]' MIN = '(80.666667-inf)' DEWP = '(-inf-64.566667]'
WDSP = '(-inf-12.2] ⇒ POST_MAX = nor
	5	0.95447

MAX = '(104.6-inf)' MIN = '(71.733333-80.666667]' DEWP = '(72.533333-inf)' ⇒ POST_MAX = high	9	0.78364

MAX = '(104.6-inf)' MIN = '(80.666667-inf)' DEWP = '(72.533333-inf)'
WDSP = '(-inf-12.2]' ⇒ POST_MAX = high
	27	0.7781

TABLE 13

Generated association rules for 48 hours before hot day prediction during summer months with support and confidence values

Association Rule $(A \Rightarrow B)$	Support $(A \cup B)$	Confidence P(B/A)
MAX = '(104.6-inf)' MIN = '(80.666667-inf)' DEWP = '(64.566667-72.533333]' WDSP = '(12.2-3.5]' ⇒ POST_MIN = high	12	0.97741
MAX = '(90.8-104.6]' MIN = '(80.666667-inf)' DEWP = '(72.533333-inf)' VISIB = '(-inf-6.766667]' WDSP = '(-inf-12.2]' 296 ⇒ POST_MIN = nor	296	0.83235
MAX = '(90.8-104.6]' MIN = '(80.666667-inf)' DEWP = '(64.566667-72.533333]' WDSP = '(-inf-12.2]' 236 ⇒ POST_MIN = nor	236	0.7757
MAX = '(104.6-inf)' WDSP = '(12.2-23.5]' 22 ⇒ POST_MIN = high	22	0.69106
MAX = '(104.6-inf)' MIN = '(80.666667-inf)' DEWP = '(72.533333-inf)' WDSP = '(-inf-12.2]' 27 ⇒ POST_MIN = high	27	0.60039

TABLE 14

Statistical summary of cold day and hot day prediction

Measures	Cold day prediction		Hot day prediction	
	24 hour before	48 hour before	24 hour before	48 hour before
Total number of instances	2304	2304	2374	2373
Correctly classified instances	1827	1837	2093	2075
Incorrectly classified instances	477	467	281	298
Correctly classified in %	79.31	79.73	88.1634	87.4421
Incorrectly classified in %	20.69	20.27	11.8366	12.5579
Mean absolute error	0.2527	0.2832	0.1878	0.1991
Root mean squared error	0.3584	0.3697	0.3077	0.3179

(*iii*) Perform classification and obtain performance metrics.

(*iv*) Select the next partition as testing and use the rest as training data.

(*v*) Repeat classification until each partition has been used as the test set.

(*vi*) Calculate an average performance from the individual experiments.

The experience of many machine learning experiments suggest that using 10 partitions (tenfold cross-validation) often yields the same error rate as if the entire data set had been used for training. By using supplied test set method, forty-five years (1961-2005) of dataset is used as training set and remaining individual years 2006, 2007, 2008, 2009 and 2010 are used as testing set respectively. First the method is used to predict the wet and dry days. Then it is tested to predict the cold day of winter and hot days of summer.

Validation is done to find out the reliability of the generated results and to show whether they can be used in real time for the prediction of rainfall using the mining approach. Validation have been done through K* methodology (Cleary and Trigg, 1995). For predicting rain occurrences, the statistical summary on SW and NE monsoon is shown in Table 5, the confusion matrix for positive and negative tuples of precipitation prediction describes in Table 6, the values of confusion matrix is shown in Table 7 and the detailed accuracy by class for precipitation prediction on SW and NE monsoon is described in Table 8. The validation results are shown in Table 9 using supplied test set method for the years 2006, 2007, 2008, 2009 and 2010 and these results are reasonable accurate. It is seen that the percentage of success in forecast is very satisfactory both in southwest and northeast monsoon seasons.

Next the methodology was tried to predict the very cold winter days and very hot summer days. The 20 °C (68 °F) was tacitly taken as threshold value for minimum temperature prediction (*i.e.*, Cool day prediction) during

January and February of winter months and 38 °C (100.4 °F) as threshold for maximum temperature prediction (*i.e*., hot day prediction) during April and May of summer months. The cold day is considered when winter minimum is below the threshold value (*i.e*., 68 °F) and the hot day is considered when the summer maximum is above the threshold (*i.e*., 100.4 °F) were predicted. The mean daily maximum temperature, mean daily minimum temperature, wind speed, dew point and visibility are considered for the occurrence of the cold day and hot day prediction. A total of 2374 instances were used for hot day prediction after filtering operations were carried out, in this a total of 310 instances were the above threshold value 38 °C (100.4 °F). 2304 instances were used for analysis for cold day prediction, in this 481 instances were below threshold value 20 °C (68 °F). The weather patterns have been generated for the occurrence of the cold day during winter months using the above weather parameters. The correlations and relationships of cold day and hot day for 24 hours ahead and 48 hours ahead are shown in Tables 10-13. The statistical validation of cold and hot day prediction using 10-fold cross validation method with the predictor accuracy measures is shown in Table 14. The loss functions absolute error and mean squared error measure the error between the actual class and the predicted class. Here the weather patterns are also very encouraging.

5. Conclusions

A forecast for predicting wet and dry days during monsoon seasons as well as the cold day during winter and hot day in summer based on data mining technique has been proposed for a sample inland station, Trichirappalli of South India. The results are encouraging and interesting. The main advantage of this model is that we need '*in-situ*' data only instead of data at different stations covering the region.

References

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 RJ 9839.

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*, Ch. 12, AAAI/MIT Press.

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.

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.

Dhar, O. N. and Rakheja, R. R., 1983, "Foreshadowing northeast monsoon rainfall over Tamil Nadu, India", *Mon. Wea. Rev*., **111**, 109-112.

Kripalani, R. H. and Kumar, Pankaj, 2004, "Northeast monsoon rainfall variability over south peninsular India *vis-à-vis* India Ocean dipole mode", *International Journal of Climatology*, **24**, 1267-1282.

Kumar, Pankaj, Rupakumar, K., Rajeevan, M. and Sahai, A. K., 2007, "On the recent strengthening of the relationship between ENSO and Northeast monsoon rainfall over South Asia", *Climate Dynamics*, **28**, 649-660.

Meganathan, S., Sivaramakrishnan, T. R. and Chandrasekhara Rao, K., 2009, "OLAP operations on the multidimensional climate data model: A theoretical approach", *Acta Ciencia Indica*, **35**, M, 4, 1233-1237.

Mohanty, U. C., Ravi, N. and Madan, O. P., 2001, "Forecasting precipitation over Delhi during south-west monsoon season", *Meteorol. Appl*., **8**, 11-21.

Mukherjee, A. K. and Sivaramakrishnan, T. R., 1977, "Surface wind and sea waves in a hurricane field", *Nature*, **207**, 237-277.

Nayagam Lorna, R., 2009, "Variability and teleconnectivity of northeast monsoon rainfall over India", *Journal of Global and Planetary Change*, **69**, 225-231.

Raghavan, S., Sivarmakrishnan, T. R. and Ramakrishnan, B., 1983, "Size distribution of radar echo as an indicator of growth mechanism in monsoon clouds around Madras", *Journal of Atmos. Sciences,* **40**, 428-434.

Raghavan, S., Sivararamakrishnan, T. R., Premkumar, S. W. and Ramakrishnan, B., 1987, "Radar reflectivity - rainfall relationship for southwest monsoon over Madras area", *Mausam*, **38**, 3, 335-340.

Raj, Y. E. A., 1996, "Inter and intra-seasonal variation of thermodynamic parameters of the atmosphere over coastal Tamil Nadu during northeast monsoon", *Mausam*, **47**, 3, 259-268.

Rao, Y. P., 1976, "Southwest monsoon", *Met. Monograph*, IMD, Monograph No. 01/76 on Monsoon, 354-364.

Rao, K. V., 1963, "A Study of the Indian northeast monsoon season", Indian Journal of Meteorology, *Hydrology and Geophysics*, **14**, 143-155.

Seetharam, K., 2009, "Arima Model of Rainfall prediction over Gangtok (Sikkim)", *Mausam*, **60**, 3, 361-367.

Shukla, J. and Mooley, D. A., 1987, "Empirical Prediction of summer monsoon rainfall over India", *Mon. Wea. Rev*., **115**, 695-703.

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.

Sivaramakrishnan, T. R. and Meganathan, S., 2012a, "Pattern visualization on meteorological data for rainfall prediction model", *Journal of Theoretical and Applied Information Technology*, **35**, 2, 173-177.

Sivaramakrishnan, T. R. and Meganathan, S., 2012b, "Data mining as tool for precipitation prediction", *Archives Des Sciences*, **65**, 3, p8.

Sivaramakrishnan, T. R. and Meganathan, S., 2012c, "Point rainfall prediction using data mining technique", *Res. Journal of App. Sciences, Engineering and Technology*, **4,** 13, 1899-1902.

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.

Sivaramakrishnan, T. R. and Sreedharan, S., 1988, "Annual Rainfall over Tamil Nadu", Hydrology Journal of IAH, **11**, 2, p20.

Sivaramakrishnan, T. R., 1988, "Rainfall characteristics of Sriharikota", ISRO Scientific Report No. 05-026-88.

Sivaramakrishnan, T. R. and Prasad, J. R., 1999, "Some aerological observations during the passage of a cyclone", *Mausam*, **50**, 215-216.

Sivaramakrishnan, T. R. and Sridharan, S., 1989, "Wind observation from Bay cyclone of November 1984", *Mausam*, **40**, 344-345.

Sridharan, S. and Muthuchami, A., 1990, "Northeast monsoon rainfall in relation to El Nino, QBO and Atlantic hurricane frequency", *Vayumandal*, **20**, 108-111.

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.