Research Journal of Applied Sciences, Engineering and Technology 5(14): 3692-3696, 2013
ISSN: 2040-7459; e-ISSN: 2040-7467

Submitted: July 05, 2012 Accepted: July 31, 2012 Published: April 20, 2013

# Association Rule Mining and Classifier Approach for 48-Hour Rainfall Prediction Over Cuddalore Station of East Coast of India

[1]S. Meganathan and [2]T.R. Sivaramakrishnan
[1]Department of CSE, SASTRA University, Kumbakonam,
[2]School of EEE, SASTRA University, Thanjavur, Tamilnadu, India

**Abstract:** The methodology of data mining techniques has been presented for the rain forecasting models for the Cuddalore (11°43′ N/79°49′ E) station of Tamilnadu in East Coast of India. Data mining approaches like classification and association mining was applied to generate results for rain prediction before 48 hour of the actual occurrence of the rain. The objective of this study is to demonstrate what relationship models are there between various atmospheric variables and to interconnect these variables according to the pattern obtained out of data mining technique. Using this approach rainfall estimates can be obtained to support the decisions to launch cloud-seeding operations. There are 3 main parts in this study. First, the obtained raw data was filtered using discretization approach based on the best fit ranges. Then, association mining has been performed on it using Predictive Apriori algorithm. Thirdly, the data has been validated using $K^*$ classifier approach. Results show that the overall classification accuracy of the data mining technique is satisfactory.

**Keywords:** Apriori algorithm, association mining, classification, $K^*$ algorithm, rainfall prediction

## INTRODUCTION

Operational forecasters base weather prediction mainly on numerical and statistical models apart from their synoptic assessment of the dynamic atmospheric system. The classical approaches attempt to model the thermo dynamic systems for grid-point prediction based on boundary conditions. Since many meteorological elements are correlated in nature, we study the meteorological data using association rule techniques, developed by the database and database mining communities and rainfall prediction is attempted using these association rules. The monsoon could be considered in a broader sense as large scale sea breeze resulting from land sea thermal contrast. Monsoon provides life giving rain and shapes the lives of millions of population and livestock in India. The performance of monsoon has extraordinary implications on the economy of tropical countries. The monsoon processes are subtle and interactive involving many components of the climate system, so much so that understanding the underlying physical processes, monitoring and predicting the behavior of monsoon continue to be challenging areas of atmospheric scientists.

One of the monsoon systems prevailing over the Indian Subcontinent is the winter Northeast (NE) monsoon, the duration of which is October to December. The northeast monsoon (Sivaramakrishnan, 1989) is well defined over coastal TamilNadu of India with some stations receiving more than 100 cm of normal seasonal rainfall. The seasonal monsoon rainfall (Raj, 1996) manifest high variability and the inter-seasonal variations are characterized by the occurrence of years of large scale droughts and large scale floods.

The NE monsoon season is also known for occurrence of intense cyclonic storms over the Bay of Bengal and some of these cyclones have attained very high intensity and has caused extensive destructions over coastal and inland regions. In order to ensure that all the relevant data are utilized by the data mining techniques, it is important to make use of micro-station data analysis.

## LITRATURE REVIEW

Several studies have been made regarding NE monsoon as recorded by several meteorologists (Duraiswamy, 1946; Nayagam, 2009; Shukla and Mooley, 1987; Shukla and Pavolino, 1983). For Tamil Nadu and East Coastal of India, the NE monsoon is more relevant as that is the main rainy season. Some works are available regarding rain assessment and forecast for NE monsoon also (Balachandran *et al*., 2006; Chew *et al*., 1998; Dhar and Rakheja, 1983; Kripalani and Pankaj, 2004; Pankaj *et al*., 2007; Lau, 1992; Raman, 2001; Rao, 1963; Sivaramakrishnan, 1989; Sivaramakrishnan *et al*., 2011a; Sivararamakrishnan and Sridharan, 1987; Sridharan and Muthuchami, 1990; Zubair and Ropelewski, 2006) to discuss the nature and prediction of rainfall for India

**Corresponding Author:** S. Meganathan, Department of CSE, SASTRA University, Kumbakonam, Tamilnadu, India

Table 1: Nominal values for atmospheric parameters

| Weather parameter | Nominal variable | Nominal values |
|---|---|---|
| Temperature | $T_L$ | <76.7 |
| (Fahrenheit) | $T_M$ | 76.7 - 81.4 |
| | $T_H$ | >81.4 |
| Dew point (Fahrenheit) | $D_L$ | <67 |
| | $D_M$ | 67 - 73 |
| | $D_H$ | >73 |
| Wind speed (Knots) | $W_L$ | <5.2 |
| | $W_M$ | 5.2 - 10.3 |
| | $W_H$ | > 10.3 |
| Visibility (Miles) | $V_L$ | <4 |
| | $V_M$ | 4 - 7 |
| | $V_H$ | > 7 |
| Precipitation (Inches) | YES | >0 |
| | NO | = 0 |

based on conventional statistics. A few isolated attempts in India (Sivaramakrishnan *et al*., 1983; Mohanty, 1994; Seetharam, 2009) are available to study the rainfall over single stations. Recently the authors (Meganathan *et al*., 2009) have attempted to analysis the weather data using on-line analytical operations on multidimensional climate data model. But all of them use conventional synoptic correlations and they are for a country or region as a whole. Hence at sub regional level where there is fairly a uniform terrain; study has to be conducted with latest tools and methodologies to predict the occurrence of the rainfall 48 hours ahead from the earlier data during the NE monsoon period. In this study, we propose a method capable of doing association rule mining (Agrawal and Srikant, 1994) on micro-station atmospheric data for a sample coastal station in east coast of India.

**Data used:** Cuddalore (Latitude 11°43' N / Longitude 79°49' E) is a coastal station in TamilNadu located in the east coast of India. This is taken as a test site. This observatory is maintained by India meteorological department and data pertaining to 1961-2010 were used for analysis. For the atmospheric parameters temperature, dew point, wind speed, visibility and precipitation (rainfall) were considered for analysis.

## METHODOLOGY

**Data preparation:** This study uses the climate data set of Cuddalore station of coastal TamiNadu of Cauvery delta basin during the period of 1961 to 2010. Our data set consists of five atmospheric variables including temperature, dew point, wind speed, visibility and rainfall. The data set that we extracted consists of the prevailing atmospheric situations 48 hours before the actual occurrence of the rain during the North East monsoon months of October, November and December. 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 3039 instances were present for analysis. The discretization algorithm produced various best-fit ranges for the five atmospheric conditions we used in analysis (Table 1).

**Association rule mining for prediction:** The problem of mining association rule was first introduced in the last decade by database communities (Agrawal *et al*., 1993, 1995; Bayardo and Agrawal, 1999; Sarawagi *et al*., 2000). Recently the authors (Sivaramakrishnan and Meganathan, 2011b, c) have reported the suitability of association rule approach for point rainfall prediction 24 h 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, as users are often more concerned about the climate along a time dimension like.

**$R_2$:** If the wind direction is east and the weather is warm, then it keeps warm for the next 48 h. For association rules mining from the filtered dataset, we used Predictive Apriori Algorithm. 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.

**Classification:** It is a form of data analysis that can be used to extract models describing important class to predict future data trends. It predicts on categorical labels. 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 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.

**Validation methods:** 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.

Table 2: Generated association rules with support and confidence values

| Association rule ($A \Rightarrow B$) | Support (A $\cup$ B) | Confidence P(B/A) |
|---|---|---|
| TEMP = '(-inf-76.733333]' DEWP='(73.233333-inf)' WIND = '(5.2-10.3]' 11 ==> PRCP = yes 11 | 11 | 0.98842 |
| TEMP = '(-inf-76.733333]' WIND ='(10.3-inf)' 4 ==> PRCP = yes 4 | 4 | 0.94277 |
| TEMP = '(76.733333-81.466667]' VISIB = '(7.1-inf)' 4 ==> PRCP = yes 4 | 4 | 0.94277 |
| DEWP = '(73.233333-inf)' VISIB = '(7.1-inf)' 4 ==> PRCP = yes 4 | 4 | 0.94277 |
| TEMP = '(-inf-76.733333]' DEWP = '(73.233333-inf)' VISIB = '(-inf-4.1]' 62 ==> PRCP = yes 55 | 62 | 0.85516 |
| TEMP = '(76.733333-81.466667]' DEWP = '(73.233333-inf)' WIND = '(-inf-5.2]' 275 ==> PRCP = yes 193 | 275 | 0.69328 |
| TEMP = '(76.733333-81.466667]' DEWP = '(73.233333-inf)' VISIB = '(-inf-4.1]' 209 ==> PRCP = yes 147 | 209 | 0.69252 |

**Cross validation:** 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:

- Separate data in to fixed number of partitions (or folds)
- Select the first fold for testing, whilst the remaining folds are used for training.
- Perform classification and obtain performance metrics
- Select the next partition as testing and use the rest as training data
- Repeat classification until each partition has been used as the test set
- 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.

**Percentage split method:** In Percentage split, the process holds out a certain percentage of the data for testing whereas the remaining is used for training the data set. In this validation method, 66.66% has been taken for training and the remaining 33.33% has been taken for testing from the extracted data set.

**Supplied test set method:** In this 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.

## RESULTS AND DISCUSSION

**Generated association rules:** Predictive mining is a task that it performs inference on the current data in order to make a prediction. Here the climate parameters such as rainfall, 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 also contain B. This is taken to be the conditional probability, P (B|A). That is:

Support ($A \Rightarrow B$) = P (A $\cup$ B)

$$Confidence(A \Rightarrow B) = \frac{\sup port\_count(A \cup B)}{\sup port\_count(A)}$$

Recently association rule mining was successfully verified for prediction of rainfall just 24 h before (Meganathan *et al*., 2009). 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. Each and every association rule consists with a support and confidence value that determines the credibility of the rule.

**Validation:** 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 correlation coefficient of the sample instances are obtained and the machine learning algorithm K* achieves an accuracy of 75.06% using cross-validation method (Table 3) and 75.70% using percentage split method (66.66% or training,

Table 3: Overall classification accuracy using 10-fold cross validation

| Stratified cross-validation | | |
|---|---|---|
| Correctly Classified Instances | 2281 | 75.06% |
| Incorrectly Classified Instances | 758 | 24.94% |

Table 4: Confusion matrix of 10-fold cross validation

| Class label | a | b |
|---|---|---|
| a = no | 2253 | 20 |
| b = yes | 738 | 28 |

Table 5: Overall classification accuracy (correlation coefficient) using percentage split validation

| Correlation coefficient of sample instances | | |
|---|---|---|
| Correctly Classified Instances | 782 | 75.70% |
| Incorrectly Classified Instances | 251 | 24.30% |

Table 6: Confusion matrix of percentage split validation

| Class label | a | b |
|---|---|---|
| a = no | 770 | 10 |
| b = yes | 241 | 12 |

Table 7: Overall classification accuracy using supplied test set validation

| Testing year | Correctly classified instances (%) | Incorrectly classified instances (%) |
|---|---|---|
| 2006 | 94.6429 | 5.3571 |
| 2007 | 98.3871 | 1.6129 |
| 2008 | 100.0000 | 0.0000 |
| 2009 | 98.1481 | 1.8519 |
| 2010 | 100.0000 | 0.0000 |

remainder for testing) (Table 5). The confusion matrix also obtained for the above method is shown in (Table 4 and 6). The validation results are shown in (Table 7) using supplied test set method and these results are reasonable accurate.

## CONCLUSION

For rainfall prediction, we have used association rule mining and instance based classifier was applied to predict rainfall with class labels "yes" for occurrence of rainfall and "no" for non occurrence of rainfall. Results show that, association rule mining forecast are reasonably accurate and can be used for predicting the occurrence of rainfall 48 h ahead. In this study we have used the K* classification approach for validating the results obtained. As evidenced in our results, the methodology can be used to facilitate the monitoring of weather conditions and predict the rain occurrence over the Cauvery delta region before the 48 h of its occurrence.

## ACKNOWLEDGMENT

The authors wish to thank National Climatic Data Center (NCDC), Asheville, North Carolina for applying data.

## REFERENCES

Agrawal, R. and R. Srikant, 1994. Fast algorithms for mining association rules. Proceeding of the Twentieth International Conference on Very Large Databases, Santiago, Chile, Sept., Expanded Version Available as IBM Research Report RJ9839.

Agrawal, R., T. Imielinski and A. Swami, 1993. Mining associations between sets of items in massive databases. Proceeding of the ACM-SIGMOD International Conference on Management of Data, Washington D.C.

Agrawal, R., H. Mannila, R. Srikant, H. Toivonen and A.I. Verkamo, 1995. Fast Discovery of Association Rules. Advances in Knowledge Discovery and Data Mining, Chapter 12, AAAI/MIT Press.

Bayardo Jr, R.J. and R. Agrawal, 1999. Mining the most interesting rules. Proceeding of the 5th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.

Balachandran, S., R. Asokan and S. Sridaran, 2006. Global surface temperature in relation to northeast monsoon rainfall over Tamil Nadu. J. Earth Syst. Sci., 115(3): 349-362.

Chew, F.H.S., T.C. Piechota, J.A. Dracup and T.A. Mcmohan, 1998. Elnino/ southern oscillation and rainfall variations. J. Hydrol., 204(1-4): 138-149.

Cleary, J.G. and L.E. Trigg, 1995. An Instance-based learner using an entropic distance measure. Proceedings of the Twelfth International Conference on Machine Learning, San Francisco, Morgan Kaufmann, pp: 108-114.

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

Duraiswamy, I.V., 1946. Scientific Note no: 98. India Meteorological Department, pp: 147.

Kripalani, R.H. and K. Pankaj, 2004. Northeast monsoon rainfall variability over south peninsular india vis-à-vis indian ocean dipole mode. Int. J. Climatol., 24: 1267-1282.

Lau, K.M., 1992. East asian summer monsoon rainfall variability and climate teleconnections. J. Meteorol. Soc. Japan, 70: 211.

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

Mohanty, V.C., 1994. Forecast of precipitation over Delhi during SW Monsoon. Mausam, 45: 87.

Nayagam, L.R., 2009. Variability and teleconnectivity of northeast monsoon rainfall over India. J. Global Planet. Change, 69: 225-231.

Pankaj, K., R. Kumar, M. Rajeevan and A.K. Sahai, 2007. On the recent strengthening of the relationship between ENSO and northeast monsoon rainfall over South Asia. Clim Dyn., 28: 649-660.

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.

Raman, K., 2001. The case for probabilistic forecast in hydrology. J. Hydrol., 249: (1-4): 2-9.

Rao, K.V., 1963. A Study of the indian northeast monsoon season. Indian J. Meteorol. Hydrol. Geophys., 14: 143-155.

Sarawagi, S., S. Thomas and R. Agrawal, 2000. Integrating association rule mining with databases: Alternatives and implications. Data Min. Knowl. Dis. J., 4(2-3): 89-125.

Seetharam, K., 2009. Arima Model of Rainfall prediction over Gangtok. Mausam, 60: 361.

Shukla, J. and D.A. Pavolino, 1983. Southern oscillation and long range forecast of summer monsoon rainfall in india. Monthly Weather Rev., 111: 1830.

Shukla, J. and D.A. Mooley, 1987. Empirical prediction of summer monsoon rainfall in India. Monthly Weather Rev., pp: 695-703.

Sivaramakrishnan, T.R., 1989. Annual rainfall over Tamil Nadu. Hydrol. J., IAH, pp: 20.

Sivaramakrishnan, T.R., *et al*., 1983. A study of rainfall over Madras. Vayumandal, pp: 69.

Sivararamakrishnan, T.R. and S. Sridharan, 1987. Occurrence of heavy rain episodes over Madras. Proceedings of National symposium on Hydrology, NIH, Roorkee, pp: VI 54.

Sivaramakrishnan, T.R., S. Meganathan and P. Sibi, 2011a. An Analysis of Northeast Monsoon Rainfall for the Cauvery Delta of Tamil Nadu. Extended Abstracts, Seminar on Indian Northeast Monsoon-Recent Advances and Evolving Concepts, Indian Meteorological Society, pp: 105-107.

Sivaramakrishnan, T.R. and S. Meganathan, 2011b: Point rainfall estimation using association rule mining-A case study. Proceedings of the National Conference on Environmental Science and Technologies for Sustainable Development, Chennai, pp: 81-85.

Sivaramakrishnan, T.R. and S. Meganathan, 2011c. Association rule mining and classifier approach for quantitative spot rainfall prediction. J. Theoret. Appl. Inform. Technol., 34(2): 173-177.

Sridharan, S. and A. Muthuchami, 1990. Northeast monsoon rainfall in relation to El NLino: QBO and Atlantic hurricane frequency. Vayumandal, 20(3-4): 1.08-1.11.

Zubair, L. and Ropelewski, 2006. The strengthening relationship of ENSO and the north east monsoon rainfall over Sri Lanka and Southern India. J. Climate, 19(8): 1567-1575.