Journal of Theoretical and Applied Information Technology
31st December 2011. Vol. 34 No.2

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

ASSOCIATION RULE MINING AND CLASSIFIER APPROACH FOR QUANTITATIVE SPOT RAINFALL PREDICTION

T.R. SIVARAMAKRISHNAN[1], S. MEGANATHAN[2]
[1]School of EEE, SASTRA University, Thanjavur, 613 401
[2]Department of CSE, SASTRA University, Kumbakonam, 612 001
E-mail: trsivaraman@yahoo.co.in, meganathan@src.sastra.edu

ABSTRACT

Rainfall prediction is usually done for a region but spot quantitative precipitation forecast is required for individual township, harbours and stations with vital installation. A methodology using data mining technique has been tried for a coastal station, Cuddalore in East Coast of India and the results are presented here. The method gives good result for the prediction of daily rainfall 24 hours ahead. There are three main parts in this work. First, the obtained raw data was filtered using discretization approach based on the best fit ranges. Then, association mining has been performed on dataset 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: *Predictive Apriori Algorithm, K^* Algorithm, Classification, Association Mining, Rainfall Prediction*

1. INTRODUCTION

Rainfall prediction is an integral part of meteorology. But operational forecasts mostly contain forecast for the region in general or a country as a whole, whether it is seasonal forecast or 24 hourly rainfall forecast. Many methods statistical, empirical and synoptic are available based on models [5], [7], [8], [10], [13]-[15]. But of late forecast for spot quantitative precipitation is gaining importance. There are many vital for individual townships, metropolis, harbours etc. A few attempts in India by Sivaramakrishnan [16]-[17], Mohanty [9] and Seetharam [12] are available. All of them use conventional statistical methods or the synoptic correlation. In this paper, a methodology of data mining technique is used for rainfall prediction, perhaps for the first time.

One of the monsoon systems prevailing over the Indian Subcontinent is the winter northeast (NE) monsoon whose period is October to December. The NE monsoon is well defined over coastal TamilNadu with some stations receiving more than 100cm of normal seasonal rainfall. The seasonal monsoon rainfall exhibit 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 have 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. Therefore, we propose a method capable of doing association mining on micro-station atmospheric data.

2. 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 since long 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.

3. METHODOLOGY

3.1 DATA PREPARATION

Many meteorological parameters are correlated in nature as they are interdependent in deciding the atmospheric dynamics. For rainfall, presence of moisture is must. Atmospheric humidity is

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

indicated by dew point. Temperature causes evaporation for adding moisture. The wind can mix the air mass causing the moisture variation. Visibility depends on aerosols which act as nuclei of condensation for the moisture to condense. Hence the parameters considered are temperature, dew point, wind speed and visibility. The data set extracted consists of prevailing atmospheric situation 24 hours before the actual occurrence of rainfall. 2999 instances of the period were present for analyzing.

Data preprocessing steps were applied on the new set of seasonal data and they were converted to nominal values by applying filters using unsupervised attribute of discretization algorithm. After the operations were carried, a total of 2999 instances were present for analysis. The discretization algorithm produced various best fit ranges [Table 1] for the five atmospheric conditions which we used in analysis.

Table 1: *Nominal Values For Atmospheric Parameters*

TEMPERATURE (Fahrenheit)	T_L	<76
	T_M	76 – 81
	T_H	>81
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

3.2 ASSOCIATION MINING FOR PREDICTION

The problem of mining association rules was first introduced by Agrawal et al [1]. The most often cited application of association rules is market basket analysis using transaction databases from supermarkets. The database contains sales transaction records, each of which details the items bought by a customer in the transaction. Mining association rules is the process of discovering knowledge such as

$Rule_1$: 80% of customers bought bread also bought jam

This $Rule_1$ can be expressed as "bread jam (25%, 80%) ", where 80% is the confidence level of the rule and 25% is the support level of the rule, indicating how frequently the customers bought both bread and jam. The discovered knowledge can then be used in store floor planning, sales promotions, etc.

When we apply the above association rule concept for analysis the meteorological data, with each record listing various atmospheric observations including wind direction, wind speed, temperature, relative humidity, rainfall and atmospheric pressure taken at a certain time in certain area we can find association rules like

$Rule_2$: If the humidity is medium wet, then there is no rain in the same area at the same time.

Although rule $Rule_2$ reflects some relationships among the meteorological elements, its role in weather prediction is inadequate, as users are often more concerned about the weather along a time dimension like.

$Rule_3$: If the wind direction is east and the weather is warm, then it keeps warm for the next 24 hour.

The traditional association rules are intra-transactional since they only capture associations among items within the same transactions, where the notion of the transaction could be the items bought by the customer, the atmospheric events that happened at the same time, and so on. However, an inter-transactional association rule can represent not only the association of items within transactions, but also the association of items among different transactions along certain dimensions like the next 6 hour, the next day, etc.

For association rules mining from the filtered dataset, we use predictive Apriori algorithm [2] for finding the hidden relationship between various atmospheric parameters. The basic property of Apriori is that all non empty subsets of a frequent item set must be frequent. A frequent item set must be frequent in connection with the above the algorithm searches with an increasing support threshold for the best 'N' rules concerning a support-based corrected confidence value.

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

3.3. CLASSIFICATION

Classification 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 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 the 24 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.

3.4. VALIDATION METHODS

Validation for our model has been done using the cross-validation and split percentage method. The basic notion of those methods has been described here.

3.4.1. 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 which often occurs is the lack of readily available training or testing data. These instances must be pre-classified which is typically time-consuming. A nice 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, while the remaining folds are used for training.

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.

3.4.2. Percentage Split Method

In Percentage split, the process hold out a certain percentage of the data for testing whereas the remaining are used for training the data set. In this validation method, two third of data has been taken for training and the remaining has been taken for testing from the extracted data set.

3.4.3. 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.

4. RESULTS AND DISCUSSION

Predictive mining is a task that it performs inference on the current data in order to make a prediction. Here the weather parameters such as rainfall, dew point, visibility, wind speed and precipitation are taken to analyze 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 $A \Rightarrow B$ (i.e., the union of sets A and B). This is taken to be the probability, $P(A \Rightarrow B)$. 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,

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

$$Confidence\ (A \Rightarrow B) = support\text{-}count\ (A \cup B) / support\text{-}count(A)$$

The predictive Apriori algorithm shows the best association rules. Some of the best rules which have been predicted from the given dataset is shown in Table 2. Each and every association rule will have a support and confidence value which determines the credibility of the rule.

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

Table 2: Generated Association Rules With Support And Confidence Value

ASSOCIATION RULE (*A* ⇒ *B*)	SUPPORT (A ∪ B)	CONFIDENCE (A ∪ 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
TEMP='(-inf-76.733333]' DEWP='(73.233333-inf)' VISIB='(-inf-4.1]' 62 ==> PRCP=yes 55	55	0.85516
TEMP='(-inf-76.733333]' VISIB='(-inf-4.1]' WIND='(-inf-5.2]' 76 ==> PRCP=yes 58	58	0.73745
TEMP='(76.733333-81.466667]' DEWP='(73.233333-inf)' WIND='(-inf-5.2]' 275 ==> PRCP=yes 193	193	0.69328
TEMP='(76.733333-81.466667]' DEWP='(73.233333-inf)' VISIB='(-inf-4.1]' 209 ==> PRCP=yes 147	147	0.69252

4.1. 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 using 10-fold cross validation method, percentage split method and supplied test set methods. These results are shown in Table 3 , Table 4 and Table 5 respectively.

TABLE 3: *Test Mode 1: 10-Fold Cross-Validation*

Stratified cross-validation	
Correctly Classified Instances 75.7252 %	2271
Incorrectly Classified Instances 24.2748 %	728
Kappa statistic	0.0747
Mean absolute error	0.3317
Root mean squared error	0.4018
Relative absolute error	87.9537%
Root relative squared error	92.5412%

TABLE 4: *Test Mode 2: Percentage Split Method (Split 66.6% For Training, Remainder Dataset For Testing)*

Correctly Classified Instances 75.4902 %	770
Incorrectly Classified Instances 24.5098 %	250
Kappa statistic	0.0563
Mean absolute error	0.3293
Root mean squared error	0.4000
Relative absolute error	87.2509%
Root relative squared error	92.0241%
Total Number of Instances	1020

TABLE 5: *Test Mode 3: Supplied Test Set Method*

Testing Year	Correctly Classified Instances	Incorrectly Classified Instances
2006	94.6429	5.3571
2007	96.7742	3.2258
2008	100.0000	0.0000
2009	98.1481	1.8591
2010	100.0000	0.0000

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

5. CONCLUSION

The association rule mining and instance based classifier approach have been applied for rainfall analysis and prediction of rainfall 24 hours ahead has been tried for a sample station. The results are reasonably accurate. Hence the methodology may be useful for quantitative precipitation forecast 24 hours ahead for the East Coast of India.

ACKNOWLEDGMENT

The authors thank the Indian Meteorological Society, Chennai Chapter for providing the opportunity for sharing information and receiving useful comments during the presentation in seminar on Indian Northeast Monsoon – Recent Advances and Evolving Concepts INEMEREC – 2011 during February of 2011.

REFRENCES:

[1] Agrawal R., Imielinski T., Swami A.: "Mining Associations between Sets of Items in Massive Databases", *Proc. of the ACM-SIGMOD 1993 Int'l Conference on Management of Data*, Washington D.C., May 1993.

[2] Agrawal R., Srikant R.,: "Fast Algorithms for Mining Association Rules", *Proc. of the 20th Int'l Conference on Very Large Databases*, Santiago, Chile, Sept. 1994. Expanded version available as IBM Research Report RJ9839, June 1994.

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

[4] Bayardo Jr R. J. and Agrawal R., "Mining the Most Interesting Rules" In *Proc. of the 5th ACM SIGKDD Int'l Conf. on Knowledge Discovery and Data Mining*, August 1999.

[5] Chew, F H S, Piechota T C, Dracup J.A, and Mcmohan T.A., "Elnino/ Southern Oscillation and Rainfall Variations", Journal of Hydrology (1998), p 138.

[6] Jiawei Han and Micheline Kamber: "Data Mining – Concepts and Techniques", Morgan Kaufmann Publications, 2008.

[7] Lau, K. M., "East Asian Summer Monsoon Rainfall Variability and Climate Teleconnections", Journal of Meteorological Society of Japan, 70 (1992), p 211.

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

[9] Mohanty V.C., "Forecast of precipitation over Delhi during SW Monsoon", Mausam, 45 (1994), p 87.

[10] Raman Kzyszie ,"The case for probabilistic forecast in hydrology", Journal of Hydrology 249, 2001, p 2.

[11] Sarawagi S., Thomas S., Agrawal R.: "Integrating Association Rule Mining with Databases: Alternatives and Implications", *Data Mining and Knowledge Discovery Journal*, **4**(2/3), July 2000.

[12] Seetharam. K , "Arima Model of Rainfall prediction over Gangtok", Mausam, 60 (2009), p 361.

[13] Shukla, J. and Mooley, D.A. , "Empirical Prediction of Summer monsoon rainfall in India", Monthly Weather Rev. (1987), p 695.

[14] Shukla, J. and Pavolino, D.A., "Southern Oscillum and long range forecast of summer monsoon rainfall in India", Monthly Weather Rev., 111 (1983), p 1830.

[15] Sivaramakrishnan, T.R, 1989, "Annual Rainfall over Tamil Nadu", Hydrology Journal, IAH, p 20.

[16] Sivaramakrishnan, T.R, et al,. 1983,"A study of rainfall over Madras - Vayumandal, p 69.

[17] Sivararamakrishnan, T.R, and Sridharan, S., 1987, "Occurrence of heavy rain episodes over Madras, Proceedings of National symposium on Hydrology, NIH, Roorkee, P VI 54.

[18] Sivaramakrishnan, T.R, Meganathan, S., and Sibi, P., "An analysis of northeast monsoon rainfall for the Cauvery delta of Tamil Nadu", Proceedings of INEMREC – 2011, pp 105-107.