Research Journal of Applied Sciences, Engineering and Technology 4(13): 1899-1902, 2012
ISSN: 2040-7467

Submitted: February 02, 2012 Accepted: March 02, 2012 Published: July 01, 2012

Point Rainfall Prediction using Data Mining Technique

[1]T.R. Sivaramakrishnan and [2]S. Meganathan
[1]School of EEE, SASTRA University, Thanjavur, India 613 401
[2]Department of CSE, SASTRA University, Srinivasa Ramanujan Centre, Kumbakonam, India 612 001

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. With recent successful attempt for prediction of rainfall at a coastal station in east coast of India, a methodology to predict spot rainfall using association rule mining for an interior station Trichirappalli (10º48′ N/78º41′ E) of south India has been developed and the results are presented here. The data is filtered using discretization approach based on the best fit ranges and then association mining is performed on dataset using Predictive Apriori algorithm and then the data need be validated using K* classifier approach. The results show that the overall classification accuracy for occurrence and non occurrence of the rainfall on wet and dry days using the data mining technique is satisfactory.

Key words: Association rule mining, data mining, K* algorithm, predictive apriori algorithm, rainfall prediction

INTRODUCTION

Rainfall prediction is a challenging task in meteorology. Conventionally rain forecast is given for a region as a whole or a country as a whole. Many methods are available for regional rainfall forecast based on statistical models (Chew *et al*., 1998; Lau *et al*., 1992; Raman, 2001; Shukla and Pavolino, 1983; Shukla and Mooley, 1987; Sivaramakrishnan, 1989). Recently the authors Meganathan *et al*. (2009) have attempted the query processing on weather dataset using on-line analytical processing operations. However of late forecast for spot quantitative precipitation is gaining importance. This is very vital for the existing and developing individual townships, metropolis, harbours etc. A few attempts in India towards spot rain analysis and prediction are available (Balachandran *et al*., 2006; Mohanty, 1994; Seetharam, 2009; Sivaramakrishnan *et al*., 1983; Sivaramakrishnan and Sridharan, 1987; Sivaramakrishnan *et al*., 2011; Zubair and Ropelewski, 2006). But most of them use conventional statistical methods or the synoptic correlation. In this study, a methodology of data mining technique is used for rainfall prediction over an inland station, Trichirappalli.

Trichirappalli (10º48′ N/78º41′ E) is located in peninsular India. The months of October to December are the main rainy period here. Rainfall is a parameter which can show wide variation. Though synoptic systems cause the convergence of moist air leading to rain, the local topography and terrain features have also a part to play in deciding the amount of rainfall. This is responsible for the wide variation of rain amount within the region. Hence a method which concentrates on the 'insitu' meteorological parameters can prove to be a potential method for rainfall prediction. Since the various meteorological parameters are interrelated the rainfall potential can be derived from an analysis of the same.

Data used: Trichirappalli (Latitude 10º48′ N/Longitude 78º41′ E) is an inland station in Tamilnadu state of South India. This is taken as a test site. This observatory is maintained by India Meteorological Department since long and data pertaining to 1961-2009 were used for analysis. For the atmospheric parameters temperature, dew point, wind speed, visibility and precipitation (rainfall) were considered for analysis.

METHODOLOGY

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 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 of sample station Trichirappalli (Latitude 10º48′ N/Longitude 78º41′ E) of South India extracted consists of prevailing atmospheric situation 24 h before the actual

Corresponding Author: S. Meganathan, Deptartment of CSE, SASTRA University, Srinivasa Ramanujan Centre, Kumbakonam, India, 612 001.

Table 1: Nominal values for atmospheric parameters

Temperature	T_L	<75.13
(Fahrenheit)	T_M	75.13-82.5
	T_H	>82.5
Dew point		
(Fahrenheit)	D_L	<60
	D_M	60-69
	D_H	>69
Wind speed		
(Knots)	W_L	<8.5
	W_M	8.5-17
	W_H	>17
Visibility		
(Miles)	V_L	<5.1
	V_M	5.1-9.86
	V_H	>9.86
Precipitation	Yes	>0
(Inches)	No	= 0

occurrence of rainfall. 3393 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 3393 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.

Association rule mining for prediction: The problem of mining association rules was first introduced in the last decade (Agrawal *et al*., 1993; Agrawal *et al*., 1994; Agrawal *et al*., 1995; Bayardo and Agrawal, 1999; Sarawagi *et al*., 2000). Recently the authors (Sivaramakrishnan and Meganathan, 2011) 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 for analysis of the 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 observation point we can find association rules like

$Rule_1$: If the humidity is medium wet, then there is no rain in the same location at the same time.

Although rule $Rule_1$ 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

Rule2: If the wind direction is east and the weather is warm, then it keeps warm for the next 24 h.

For association rules mining from the filtered dataset, we use predictive Apriori algorithm (Agrawal *et al*., 1994) 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.

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 h 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 cross-validation and split percentage method. The basic notion of those methods has been described here.

Cross validation method: 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:

- Separate data in to fixed number of partitions (or folds)
- Select the first fold for testing, while 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.

Table 2: Generated association rules for climate dataset with support and confidence value

Association rule (A ⇒ B)	Support (A ∪ B)	Confidence (A ∪ B/A)
TEMP = '(75.133333-82.466667]' DEWP = '(60-68.9]' VISIB = '(5.133333-9.866667]' WIND = '(-inf-8.533333]' 31 ==> PRCP = yes 31	31	0.99155
TEMP = '(75.133333-82.466667]' DEWP = '(60-68.9]' WIND = '(8.533333-17.066667]' 262 ==> PRCP = yes 242	262	0.9198
TEMP = '(82.466667-inf)' DEWP = '(60-68.9]' 74 ==> PRCP = yes 69	74	0.91947
DEWP = '(60-68.9]' VISIB = '(-inf-5.133333]' WIND = '(8.533333-17.066667]' 284 ==> PRCP = yes 259	284	0.91431
VISIB = '(9.866667-inf)' WIND = '(8.533333-17.066667]' 3 ==> PRCP = yes 3	3	0.89357
TEMP = '(82.466667-inf)' WIND = '(17.066667-inf)' 2 ==> PRCP = yes 2	2	0.86124
DEWP = '(-inf-60]' VISIB = '(5.133333-9.866667]' 2 ==> PRCP = yes 2	2	0.86124
TEMP = '(-inf-75.133333]' DEWP = '(60-68.9]' WIND = '(-inf-8.533333]' 61 ==> PRCP = yes 52	61	0.84313
TEMP = '(-inf-75.133333]' DEWP = '(68.9-inf)' 99 ==> PRCP = no 81	99	0.80334
DEWP = '(68.9-inf)' VISIB = '(9.866667-inf)' 4 ==> PRCP = no 3	4	0.70693
DEWP = '(68.9-inf)' WIND = '(17.066667-inf)' 4 ==> PRCP = no 3	4	0.70693
TEMP = '(-inf-75.133333]' WIND = '(8.533333-17.066667]' 51 ==> PRCP = no 35	51	0.68675

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.

Supplied test set method: In this method, forty five years (1961-2005) of dataset is used as training set which contains 2943 instances and remaining individual years 2006, 2007, 2008 and 2009 are used as testing set respectively.

RESULTS AND DISCUSSION

Association rule extraction: 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 rule mining. The rule A⇒B holds in the transaction set D with support s, where s is the percentage of transactions in D that contain A⇒B (i.e., the union of sets A and B). This is taken to be the probability, P (A⇒B). The rule A⇒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:

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

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

The predictive Apriori algorithm shows the association rules for the occurrence and non occurrence of the rainfall with interested patterns of climate parameters on wet and dry days. Some of the best rules which have been predicted from the given dataset are shown in Table 2. Each and every association rule will have a support and confidence value which determines the utility and certainty of the association rule.

Table 3: Test mode 1-10 fold cross-validation

Stratified cross-validation	
Correctly classified instances 66.4309%	2254
Incorrectly classified instances 33.5691%	1139
Kappa statistic	0.0728
Mean absolute error	0.4197
Root mean squared error	0.4503
Relative absolute error	91.7907%
Root relative squared error	94.1908%
Total number of instances	3393

Table 4: Test mode 2-percentage split method (66.6% for training and remainder dataset for testing)

Correctly classified instances 65.1646%	752
Incorrectly classified instances 34.8354%	402
Kappa statistic	0.072
Mean absolute error	0.4269
Root mean squared error	0.458
Relative absolute error	92.9079%
Root relative squared error	94.8588%
Total number of instances	1154

Table 5: Test mode 3-supplied test set method

Testing year	Correctly classified instances (%)	Incorrectly classified instances (%)
2006	56.044	43.956
2007	69.2308	30.7692
2008	72.5275	27.4725
2009	63.3333	36.6667

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. Validations 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, 4 and 5, respectively.

CONCLUSION

The association rule mining and instance based classifier approach have been applied for rainfall analysis and prediction of rainfall 24 h ahead for a sample station in interior Tamilnadu state of South India. Association rule extraction is performed for the occurrence and non occurrence of the rainfall on wet and dry days with support and confidence values. The extracted climate

patterns are interesting since they are all easily understood, valid on test data with some degree of certainty, potentially useful and novel. The results are reasonably accurate. Hence the methodology may be useful for quantitative precipitation forecast 24 h ahead for East Tamilnadu state of India.

ACKNOWLEDGMENT

The authors wish to thank National Climatic Data Center (NCDC), Asheville, North Carolina for supplying data and thank the Indian Meteorological Society, Chennai Chapter for providing the opportunity for sharing information and receiving useful comments during the presentation in National Conference on Environmental Science and Technologies for Sustainable Development during August 2011.

REFERENCES

Agrawal, R., T. Imielinski and A. Swami, 1993. Mining Associations between Sets of Items in Massive Databases. Proceeding of the ACM-SIGMOD 1993 Int'l Conference on Management of Data, Washington D.C.

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

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.

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.

Bayardo, Jr, R.J. and R. Agrawal, 1999. Mining the Most Interesting Rules. In Proc. of the 5th ACM SIGKDD Int'l Confrence on Knowledge Discovery and Data Mining.

Chew, F.H.S., T.C. Piechota, J.A. Dracup and T.A. Mcmohan, 1998. Elnino/Southern Oscillation and Australiarr rainfall, streamflow and drought: Links and potential for forecasting. J. Hydrol., 204(1-1): 138-149.

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*, XXXVM, 4: 1233-1237.

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

Raman, K., 2001. The case for probabilistic forecast in hydrology. J. Hydrol., 249: 2.

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

Seetharam, K., 2009. Arima model of rainfall prediction over Gangtok. Mausam, 60: 361.

Shukla, J. and D.A. Mooley, 1987. Empirical prediction of Summer monsoon rainfall in India. Monthly Weather Rev., 115: 695-704.

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

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.

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

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

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

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.