Fuzzy Association Rules for Prediction of Monsoon Rainfall

C.T. Dhanya and D. Nagesh Kumar

Department of Civil Engineering, Indian Institute of Science, Bangalore, India

Abstract. A fuzzy association rule algorithm is implemented to extract the relationship between the atmospheric indices and the Indian Summer Monsoon Rainfall (ISMR). ENSO and EQWIN indices are used as the causative variables. Rules extracted are showing a negative relationship with ENSO index and a positive relationship with the EQWIN index. A fuzzy rule based prediction technique is also implemented on the same indices in order to predict the ISMR. Rules are defined using a training dataset for the period 1958 – 1999 and validated for the period 2000 – 2006. The fuzzy outputs of the defined rules are converted into crisp outputs using the weighted counting algorithm. The variability of the ISMR over the years is captured by this technique, thus proving to be efficient even when the linear statistical relationship between the indices is weak.

Key words: Fuzzy association rules, enso, eqwin, Indian summer monsoon rainfall, weighted counting algorithm

1 Introduction

The influence of various climatic and atmospheric indices on Indian summer monsoon rainfall (ISMR) has been studied intensively by various studies since the last few decades. Such studies are of great significance since majority of Indian rainfall is contributed by summer monsoon rainfall. The predictor variables used for the study of ISMR include Darwin sea-level pressure, Latitudinal position of 500 mb ridge along 75° E, Arabian sea SST, Indian Ocean SST, Quasi-Biennial Oscillation, Sea-surface temperature anomalies over different Nino regions, QBO, western Pacific region SST, eastern Indian Ocean region SST, Eurasian surface temperature and Indian Surface temperature, Equatorial East Indian Ocean sea surface temperature, Nino 3.4, Equatorial Indian Ocean Oscillation zonal wind index (EQWIN), Eurasian snow cover, NW Europe temperature. The complex relationship between these large scale circulation patterns and ISMR leads to a poor performance of the models used so far to forecast ISMR [7,12].

Among these indices, El Nino and Southern Oscillation (ENSO) and Equatorial Indian Ocean Oscillation (EQUINOO) together can explain much of the ISMR variability [5]. [7] found that a composite index which is a linear combination of the

EQWIN and the ENSO index (Nino 3.4 SST), correlates better with an excess or deficit in ISMR than either index alone. They separated the excess and deficit ISMR events by a line determined by a linear combination of the EQWIN and ENSO index.

However, since the correlation between ISMR and ENSO index and also EQUINOO index are very small (0.33 and 0.19 respectively), traditional linear statistical modelling approaches are meaningless, even at seasonal timescale [6]. In this study, a fuzzy data mining algorithm is used to generate the association rules between ENSO, EQUINOO indices and ISMR. Also, a fuzzy rule-based technique is used to quantify the relationship of ENSO and EQUINOO index with ISMR and thereby reproduce the variability of ISMR.

2 Data Used

1. ENSO index: Sea surface temperature anomaly from Nino 3.4 region (5°S–5°N, 170° E – 120° W). Monthly sea surface temperature data from Nino 3.4 region for the period, January 1958 to December 2006, are obtained from the Website of National Weather Service, Climate Prediction Centre of NOAA (http://www.cpc.noaa.gov/data/indices/).
2. EQWIN index: Negative of zonal wind anomaly over equatorial Indian Ocean region (60° – 90°E, 2.5°S – 2.5°N). Monthly surface wind data for the period, January 1958 to December 2006, are obtained from National Center for Environmental Prediction (http://www.cdc.noaa.gov/Datasets).
3. Indian rainfall: Monthly rainfall data over entire India are obtained for the period, January 1958 to December 2006, from the Website of Indian Institute of Tropical Meteorology (IITM), Pune, India (http://www.tropmet.res.in/data.html).

According to [10], ENSO and EQWIN index at a lag of one month can better separate the extreme positive and negative anomalies of monsoon rainfall. Most of the extreme positive anomalies occurred when ENSO index < EQWIN and negative anomalies occurred when ENSO index > EQWIN. Hence for the present study, standardized values of the ENSO and EQWIN index for the months May – August are considered as the antecedents. Standardized values of the Indian rainfall anomalies for the monsoon months (June – September) are selected as the consequents.

3 Fuzzy Association Rules

Association rules indicate whether or how much the values of an attribute depend on the values of the other attributes in the data set. A rule consists of a left-hand side proposition (antecedent) and a right hand side proposition (consequent). The rule states that when the antecedent occurs (is true), then the consequent also occurs (is true). The conditional probability of the occurrence of the consequent given the antecedent is referred to as the confidence of the rule. For example, if a pattern "B follows A" occurs n_1 times and the pattern "C follows B follows A" occurs n_2 times, then the association rule "whenever B follows A, C will also follow" has a confidence of (n_2/n_1). The interestingness of a rule is usually measured in terms of its confidence.

In association rule mining, subdivision of the quantitative values into crisp sets would lead to over or underestimating values near the borders. Some of the rules may have been omitted since the combination of the antecedents is not falling in the same discrete range, but they may be taking the nearby discrete states. Fuzzy sets can overcome that problem by allowing partial memberships to the different sets. [3] and [4] have shown the importance of fuzzy partition over crisp partition. Hence, a better extraction of the rules may be possible if the classification of the indices is done in a fuzzy manner and not in a crisp manner. The approach used to discover fuzzy association rules in the present study is described below:

3.1 Fuzzy Set Construction

All the three data sets are divided into five classes: (1) Lowest, (2) Low, (3) Medium, (4) High and (5) Highest. The values and overlap used for the classification are shown in figure 1.

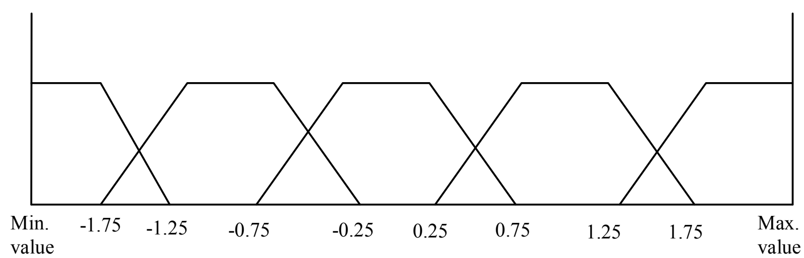

Fig. 1. Boundary values and overlap used for fuzzy set classification

The membership functions are computed based on whether it is at the upper border of a set or at the lower border.

At the upper border,

$$\mu(x) = \frac{ub(f^n) - x}{ub(f^n) - lb(f^{n+1})} \tag{1}$$

For the lower border,

$$\mu(x) = \frac{x - lb(f^n)}{ub(f^n) - lb(f^{n+1})} \tag{2}$$

where $lb(f^n)$ is the low border of a set, $ub(f^n)$ is the upper border and x is the original value of any attribute in the database.

The abbreviations used for all indices falling in different fuzzy classes are given in table 1.

Table 1. Abbreviations used for the classification

Fuzzy set	ENSO	EQWIN	ISMR
1. Min value to -1.25	Nino1	Eqwin1	Extreme drought
2. -1.75 to -0.25	Nino2	Eqwin2	Moderate drought
3. -0.75 to 0.75	Nino3	Eqwin3	Normal
4. 0.25 to 1.75	Nino4	Eqwin4	Moderate flood
5. 1.25 to Max. value	Nino5	Eqwin5	Extreme flood

3.2 Preparation of Dataset for Mining

A new data set has to be constructed from the original database based on the definition of fuzzy sets described above. For every fuzzy set defined, there is one row in the new database containing the grade of membership of the single items to the specific set. In the present study, since five fuzzy sets are defined for each of the three attributes, there will be a total of 15 columns, the first 5 columns giving the membership values of first attribute for all five fuzzy sets. Thus, the new table will only contain the membership values to these fuzzy sets.

3.3 Support and Confidence of fuzzy rules

Triangular norms are used for the calculation of quality measures, support and confidence. A triangular norm is a commutative, associative, non-decreasing function in $T:[0,1]^2 \rightarrow [0,1]$ such that $T(x,1)=x$ for all $x \in [0,1]$. The basic continuous t-norms are the minimum, the product and the Lukasiewicz t-norms.

For an association rule A→B, the support and confidence are given as:

$$\sup(A \rightarrow B) = \sum_{x,y \in D} T(A(x), B(y)) \tag{3}$$

$$conf(A \rightarrow B) = \frac{\sum_{x,y \in D} T(A(x), B(y))}{\sum_{x \in D} A(x)} \tag{4}$$

For the present study Lukasiewicz t-norm is used for the calculation of support and confidence, since it is the most popular method for calculating fuzzy operations. Hence, support and confidence can be expressed as,

$$\sup(A \rightarrow B) = \sum_{x,y \in D} \min(A(x), B(y)) \tag{5}$$

$$conf(A \rightarrow B) = \frac{\sum_{x,y \in D} \min(A(x), B(y))}{\sum_{x \in D} A(x)} \tag{6}$$

3.4 Frequent Pattern (FP) Growth Algorithm

The FP-Growth algorithm allows generating frequent itemsets and unlike the Apriori algorithm, it does not create huge amount of candidates. The data is organized in a tree form, called the Frequent Pattern Tree. The algorithm first constructs the tree out of the original data set and then grows the frequent patterns.

The data is preprocessed before applying the algorithm. The dataset is scanned for a first time to compute the support of the single items. The items that have a support value less than a user specified minimum support are discarded. The remaining items are recombined so that they appear in a decreasing order with respect to their support. An example of the dataset preprocessing [9] is given in the table 2.

Table 2. Example data base

Original database	Support	Preprocessed database
abd	sup(b) = 6	bda
bcde	sup(d) = 5	bde
bd	sup(e) = 5	bd
ade	sup(a) = 4	dea
ab	sup(c) = 2	ba
abe	Min.sup = 3	bea
cde		de
be		be

Now, for constructing the FP – tree, a scan over the database is made, adding each item to the tree. The first itemset will be the first branch of the tree. The second transaction shares a common prefix with the already existing set in the tree. In this case, the values along the path of the common prefix will be increased by one, and the remaining items will make new nodes for the tree. In the example given above, the first itemset is bda. Hence the first branch of the tree would be the items *b*, *d* and *a*. While adding the second itemset bde, a new node is introduced for *e* and also the values along the path are increased by one. The tree for the example database [9] is shown in figure 2.

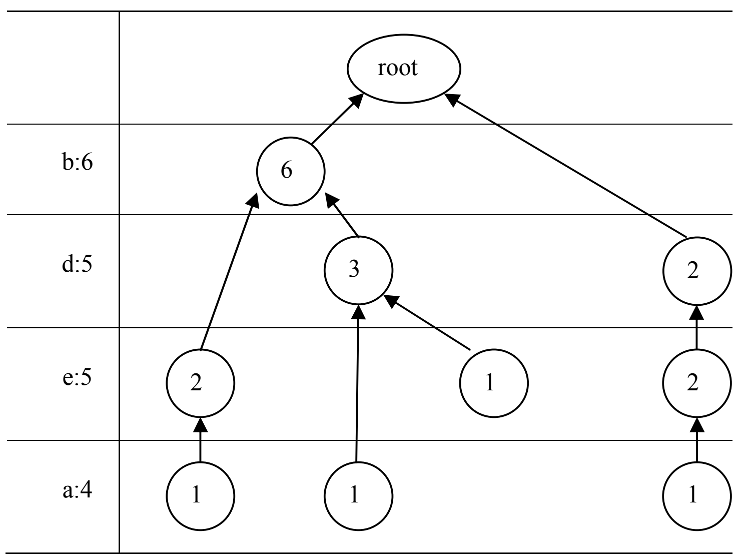

Fig. 2. FP-Tree

Each node of the FP-Tree consists of three fields [9]:
(a) *item-name*: The name of the item that the node represents is stored.
(b) *count*: The accumulated support of the node within the current path.
(c) *node-link*: Links have to be built between the nodes. It stores the ancestor of the current node, and null if there is none.

For discovering all frequent itemsets, the FP-Growth algorithm takes a look at each level of depth of the tree starting from the bottom and generating all possible itemsets that include nodes in that specific level. A detailed procedure of FP growth algorithm can be found in [8,9]. The procedure of the algorithm is given in below.

Procedure FR-Growth (*Tree,* $\boldsymbol{\alpha}$) {
1) **if** *Tree* contains a single path *P*
2) **then for each** combination ($\boldsymbol{\beta}$) of the nodes in the path *P* **do**
3) generate pattern $\beta \cup \alpha$ with support = minimum support of nodes in $\boldsymbol{\beta}$
4) ***else for each*** a_i in the header of *Tree* **do** {
5) generate pattern $\beta = a_i \cup \alpha$ with support = a_i . support;

6) construct ***β***'s conditional pattern base and then ***β***'s conditional FP_Tree, $\boldsymbol{Tree_\beta}$
7) **if** $Tree_\beta \neq \Phi$
8) **then** call FP-growth ($\boldsymbol{Tree_\beta}$, ***β)***}

}

FP tree algorithm can be used effectively for fuzzy database also. But while dealing with fuzzy database, both support of occurrences and value of membership are to be considered. The membership value is simply added to the overall count of the node.

FP algorithm can be successively used to find association rules of rare events also, by specifying the rare events as specific target events [2]. FP algorithm is particularly useful for large data bases since it scans the data only twice. However, it is equally efficient for small data bases also since it avoids the candidate generation step of Apriori algorithm. A performance study by [8] has shown that FP growth algorithm is about an order of magnitude faster than the Apriori algorithm and also faster than some recently reported new frequent-pattern mining methods, for both short and long frequent patterns.

After constructing the FP tree using the membership values, association rules are generated. The antecedent can consist of any number of items, but in the consequent, there is only one item allowed.

In the present study, ENSO and EQWIN indices are considered as the antecedents and ISMR as the consequent. Interestingness of the rules are decided based on the rule confidence. Rules for which the confidence exceeds the minimum confidence threshold are stored.

4 Fuzzy Rule Based Prediction

A fuzzy rule based modeling approach is used for the prediction of rainfall by utilizing the linkage between the ISMR and ENSO-EQWIN indices. The entire dataset from 1958 - 2006 is divided into two sets – training set (1958-19990 and validation set (2000-2006). Fuzzy rules are constructed using the training set by applying the weighted counting algorithm [1]. The algorithm consists of the following steps:

1. Classify the attributes (both antecedents and consequent) into five classes. Let the antecedent be X_1 and X_2 and the consequent be Y. The five fuzzy sets for each attributes can be expressed as $X_1^1, X_1^2, ..., X_1^5$; $X_2^1, X_2^2, ..., X_2^5$ and $Y^1, Y^2, ..., Y^5$.
2. Calculate the membership functions of all the attributes for the entire training set, thus replacing the original training database with a database of membership functions. For each attribute, find the maximum membership value at each data point. Thus, each $X_{i,j}$ (i=1,2; j=1,2,…,length of training set,n_t) data point possess a value $M_{i,j}$ which is the maximum membership value. Similarly each Y_j (j=1,2,…, n_t) possess a value $M_{o,j}$.

3. To combine the effect of all the antecedents, the degree of fulfillment (DOF) of each dataset of the training set is calculated as the product of all M_i's for each j where j=1,2,…, n_t. i.e.,

$$DOF_j = \prod_{i=1}^{2} M_{i,j} \tag{7}$$

Degree of fulfillment (DOF) indicates the degree of applicability of the rule within the system.

4. For each rule, a weight is assigned to each rule as the product of DOF and membership value of the consequent. If the rule is repeating in the database the weights are added upon. Hence, weight for k^{th} rule can be expressed as

$$wt_k = \sum_{j=1}^{n_t} .DOF_j .M_{o,j} \tag{8}$$

After scanning throughout the entire training set, all the derived rules will possess a weight.

5. Validation of the rules: Calculate the DOF of all attributes for all data points in the validation set. Identify those rules that are showing similar antecedent conditions and also having a minimum DOF value, from the derived ones.
6. Defuzzification: Each rule leads to a fuzzy response. A crisp output can be obtained from all these rules through defuzzificaiton process. In weighted counting algorithm, the center of gravity is commonly used to obtain the estimated value of the consequent variable. Hence, the estimated value can be expressed as

$$\hat{Y}_j = \frac{\sum_{k=1}^{n} .DOF_k .wt_k .B_k^{(2)}}{\sum_{k=1}^{n} .DOF_k .wt_k} \tag{9}$$

where k = number of similar rules extracted for a given antecedent combination; $B_k^{(2)}$ = most likely value of the consequent of the k^{th} rule, i.e, the value for which the membership function is one.

5 Results

5.1 Fuzzy Association Rules

FP growth algorithm is applied on the entire dataset from 1958 – 2006 to derive fuzzy rules. Rules are extracted for drought, flood and normal conditions based on the confidence measure. The rules extracted and the corresponding confidences are shown in table 3.

Table 3. Fuzzy Association Rules

Antecedent	Consequent	Confidence
Nino5, Eqwin1	Severe drought	0.95
Nino5, Eqwin1	Moderate drought	0.83
Nino5, Eqwin3	Moderate drought	0.66
Nino1,Eqwin4	Moderate flood	1.0
Nino2,Eqwin5	Moderate flood	0.83
Nino1,Eqwin1	Normal	1.0
Nino2,Eqwin1	Normal	0.79
Nino5,Eqwin5	Normal	1.0

The rules for the extremes (droughts and floods) are in concordance with the previous works done by [7] and [11] combining both ENSO and EQWIN indices for the prediction of ISMR. Although, no rules are extracted for extreme flood condition, upon reducing the minimum confidence level, it is found that a combination of Nino1 + Eqwin3 leads to an extreme flood condition; but with a very less confidence of 0.38.

5.2 Fuzzy Rule Based Prediction

Rules are defined based on the training set for the period 1958 – 1999. Validation is done for the period 2000 – 2006. Prediction is done by extracting rules similar to the antecedent combinations in the validation set. For example the antecedent combination for the first data point in the validation set is Nino2 and Eqwin4. The maximum membership values are 0.78 and 1 respectively. Now from the training database, we have extracted 7 rules which have similar or nearby antecedent combinations. The rules are given in table 4.

Table 4. Selected Rules for Predicting the First Dataset in the Validation database

Antecedent	Consequent	DOF	Weight	$B^{(2)}$
Nino2,Eqwin4	Moderate flood	0.78	1.95	1
Nino2,Eqwin4	Normal	0.81	0.98	0
Nino3, Eqwin4	Moderate flood	0.88	4.97	1
Nino3, Eqwin4	Normal	0.67	5.38	0
Nino3, Eqwin4	Moderate drought	0.57	0.76	-1
Nino3, Eqwin4	Severe flood	0.97	1.67	2

Using the formula (9), the crisp value of the consequent can be calculated as follows:

$$\hat{Y}_j = \frac{(0.78\times1.95\times1)+(0.81\times0.98\times0)+(0.88\times4.97\times1)+(0.67\times5.38\times0)+(0.57\times0.76\times-1)+(0.97\times1.67\times2)}{(0.78\times1.95)+(0.81\times0.98)+(0.88\times4.97)+(0.67\times5.38)+(0.57\times0.76)+(0.97\times1.67)} \quad \textbf{(10)}$$

$$= 0.704.$$

The predicted value is much nearer to the original value of 0.86. The consequent values are predicted for the entire validation set in a similar fashion. The predicted values for the period 2000-2006 are shown in the figure 3. The correlation coefficient between the monthly observed ISMR and the predicted ISMR is 0.63 and the root mean square error is 1.02.

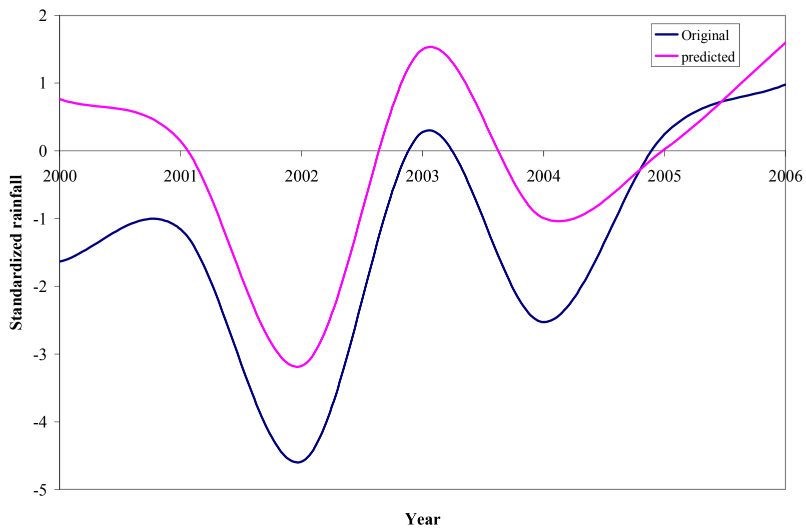

Fig. 3. Observed and predicted ISMR values for the validation period

It is evident from the results that the model performs quite well in the validation period. Since rainfall is a complex phenomena influenced by many climatic and atmospheric indices, the inclusion of more indices in the antecedent group may improve the performance of the model.

6 Conclusions

Fuzzy association rule helps in extracting the relationship between the variables, by overcoming the sharp boundary problem when mining association rules from quantitative data. The rules extracted using the algorithm give a one to one relationship of the antecedents with the consequent, without going into the complex mathematical relationships between them. The support of the rules is not taken into account, since for the extreme rainfall events, support may be very small, when compared with the normal rainfall events and thus may lead to the omission of these infrequent events.

Fuzzy rule based prediction using the weighted counting algorithm provides an excellent tool for the prediction of monthly rainfall in monsoon months using the two indices, ENSO and EQWIN. The variability of ISMR is reasonably captured using this method. This is a better substitute of the usual statistical methods, which are limited mostly by the linear statistical relationship between the indices.

References

1. Bardossy, A., Duckstein, L.: Fuzzy rule-based modeling with applications to geophysical, biological and engineering systems, CRC, New York (1999).
2. Berberidis, C., Angelis, L., Vlahavas, I.: PREVENT An algorithm for mining intertransactional patterns for the prediction of rare events, Proc. Second Starting AI Researchers' Symposium, Frontiers in Artificial Intelligence and Applications, 9 (2004).
3. Caulfield, H.J.: Fuzzy optical meteorology, IEEE Transactions on Fuzzy Systems, 4 (2) (1996).
4. Chen, Z., Chen, G.: An approach to classification based on fuzzy association rules, Advances in Intelligent Systems Research, International Conference on Intelligent Systems and Knowledge Engineering (2007).
5. Gadgil S.: The Indian Monsoon and its variability. Annual Review of Earth Planetary Science, 31, 429–467 (2003).
6. Gadgil, S., Rajeevan, M., Nanjunduah, R.: Monsoon prediction: why yet another failure, Current Science, 88(9), 1389 – 1400 (2005).
7. Gadgil, S., Vinayachandran, P. N., Francis, P. A., Gadgil, S.: Extremes of the Indian summer monsoon rainfall, ENSO and equatorial Indian Ocean oscillation, Geophys. Res. Lett., 31, doi:10.1029/2004GL019733 (2004).
8. Han, J., Pei, J., Yin, Y., Mao, R.: Mining frequent patterns without candidate generation: A frequent-pattern tree approach, Data Mining and Knowledge Discovery, 8, 53–87 (2004).
9. Han, J., Pei, J., Yin, Y.: Mining Frequent Patterns without Candidate Generation, 2000 ACM SIGMOD Intl. Conference on Management of Data, ACM Press (1999).
10. Maity, R., Nagesh Kumar, D., Nanjundiah, R.S.: Review of hydroclimatic teleconnection between hydrologic variables and large-scale atmospheric circulation indices with Indian perspective, ISH Journal of Hydraulic Engineering, 13(1), 77-92 (2007).
11. Maity, R., Nagesh Kumar, D.: Hydroclimatic association of monthly summer monsoon rainfall over India with large-scale atmospheric circulation from tropical Pacific Ocean and Indian Ocean region, Atmospheric Science Letters, 7(4), 101 – 107 (2006).
12. Rajeevan, M., Pai., D.S., Diskhit, S.K., Kelkar, R.R.: IMD's new operational models for long range forecast of southwest monsoon rainfall over India and their verification for 2003, Current Science, 86(3), 422-430 (2004).