Agris on-line Papers in Economics and Informatics

Volume IX Number 2, 2017

Statistical Feature Ranking and Fuzzy Supervised Learning Approach in Modeling Regional Rainfall Prediction Systems

M. Sudha[1], K. Subbu[2]

[1] School of Information Technology & Engineering, VIT University, India.

[2] College of Food Agricultural & Environmental Studies, Ohio State University, USA

Abstract

Rainfall prediction is an essential and challenging task in hydro-meteorology. Most of the existing weather dataset used for prediction consists of observatory record of several atmospheric parameters. Identifying the significant parameters from irrelevant and redundant parameter set for weather prediction is important because irrelevant parameters may decrease the prediction accuracy. The main intent of this research is to identify the influencing weather parameters for improving daily rainfall forecast efficiency. A parameter selection module identifies the significant parameter based on information gain based feature ranking. Fuzzy supervised learning module evaluates the performance of fuzzy classifiers before and after parameter selection. In the evaluation phase, learning techniques was analyzed in terms of Accuracy Rate (AcR), Root Mean Squared Error (RMSE) and Misclassification Rate (McR). Experimental results revealed that, parameter subset selection has significantly improved the performance of the learning techniques. The investigation results identified minimum temperature, relative humidity and evapotranspiration as influencing weather parameters for rainfall prediction. Empirical results revealed Fuzzy Unordered Rule Induction Algorithm (FURIA) as a suitable rainfall prediction approach. This fuzzy model achieved an enhanced accuracy rate of 84.10% after parameter selection with nominal misclassification rate of 0.1590%.

Keywords

Short-range rainfall prediction, statistical feature ranking, fuzzy rule induction and prediction accuracy.

Sudha, M. and Subbu, K. (2017) “Statistical Feature Ranking and Fuzzy Supervised Learning Approach in Modeling Regional Rainfall Prediction Systems", *AGRIS on-line Papers in Economics and Informatics*, Vol. 9, No. 2, pp. 117 - 126. ISSN 1804-1930. DOI 10.7160/aol.2017.090210.

Introduction

Rainfall prediction plays a vital role in most of our day to day real life activities. Especially in countries like India that depends on agricultural productivity for its economic growth need reliable weather forecasting mechanism. In India about 50% of agricultural cultivation and yield are mainly influenced by the rainfall. There exists an everlasting demand for enhanced prediction models for strategic decision support. Many of our day to day activities are influenced by that day’s weather. Therefore, rainfall prediction outcomes serve as an important factor for strategic decision support in real life activities. This analysis focuses on identifying relevant parameter for enhanced rainfall forecasting using dimension reduction approach. Dimension reduction is a challenging task in data mining and knowledge representation of high dimensional data set. It is a method of reducing the high dimensional data space to minimal dimensional space by removing irrelevant and redundant data. Parameter reduction is achieved either by feature selection or feature transformation process. Parameter selection is a method of finding the most suitable subset of the complete feature vector. It is stated that selection is achieved using statistical measures such as entropy, information gain, correlation, covariance and other data mining approaches (Ishibuchi and Nakashima, 2001, 2005). Feature transformation is the other way of reducing the data space, in this technique the features are transformed as factors representing significant features. In any feature selection technique finding the most suitable subset is a tough and exhaustive. Hence, feature selection problems are considered Nondeterministic Polynomial time (NP) hard problem (Blum and Rivest, 1992). Feature selection methods are categorized as filter (Huhn

and Hullermeier, 2009), wrapper (Nikam and Meshram, 2013) and embedded approaches.

A, information gain based feature selection technique is implemented for identifying the suitable weather parameters(Novakovic, 2009). (Siedlecki and Sklansky,1988) Automatic feature selection approach and supervised learning techniques are expected to perform better than when trained with complete feature set. In a recent trend fuzzy concepts are used in a wide range of applications such as data analytics, pattern recognition, soil evaluation and meteorology from the time of its introduction (Zadeh, 1965). This proposed approach also uses the benefits of a fuzzy based learning approach for training the system to classify with less misclassification rate. As a recent trend bio inspired techniques are used in meteorology prediction nowadays (Lee et al., 2012). Genetic algorithm based feature selection is applied for heavy rain prediction at South Korea Empirical analysis conducted on rainfall data collected for a period of 20 years, genetic algorithm based feature selection performed better than the traditional feature selection method (Seo et al., 2012).

(Liu et al., 2001) introduced a novel enhanced Naive Bayes classifier technique and explored the use of genetic algorithm for feature subset selection for classification. (Dai and Xu, 2013) described the effect of fuzzy based feature reduction approach using fuzzy gain ratio for medical dataset. The feature selection method based on the fuzzy gain ratio of fuzzy rough set theory performed better than other approaches (Maqsood et al., 2014). Sudha and Valarmathi, 2013) mentioned that a feature reduction approach based on quick reduct, entropy measure and rough set approaches have wide scope of application. (Yu, 2005) described integrated feature selection approach. The rough set feature reduction techniques, computed several reduced sets than any other approaches (Sudha and Valarmathi, 2015, 2016). Dai and (Xu, 2013) and (Blum and Rivest, 1992) described a hybrid rough fuzzy neural network model for weather forecasting.

Effect of proposed fuzzy based automated weather forecasting model using temperature to predict the daily temperature is discussed in (Al-Matarneh, 2014). The experimental results of shown that the proposed fuzzy based model enhanced accuracy rate (Maqsood et al., 2004). The performance of neural network Multi-layer Perceptron (MLP), random forest, classification and regression tree, support vector machine, and k-nearest neighbor algorithms are examined in terms of accuracy (Kusiak et al., 2014). Experimental results conveyed data mining techniques as a suitable approach to construct predictive models for normal as well time series radar data. (Zadeh, 1965) proposed hybrid intelligent systems based on rough sets, neural networks, fuzzy sets and other optimization methods. It is stated that hybrid intelligent computational approach can handle uncertain, noisy and incomplete data set. Most of the hybrid intelligent systems are cost effective solutions for various scientific applications (Li and Liu, 2005) and (Zadeh, 1965). The rainfall prediction evaluation results for Mashhad meteorology stations using Adaptive Neural Fuzzy Inference System (ANFIS) outperformed other non ANFIS models. This model considered temperature, relative humidity, cloud cover total and due point as input parameters Niksaz and Latif (2014). It is stated that, hybrid intelligent computing approaches outperform than other the traditional methods (Niksaz and Latif, 2014). As stated in (Niksaz and Latif, 2014), (Seo et al., 2014) and (Liu et al., 2001) this proposed investigation on rainfall prediction uses eight atmospheric parameters in the Coimbatore region of India. This proposed approach uses an effective feature subset of the complete feature vector and fuzzy based classifier for evaluation. It is a well-known theory, that fuzzy techniques can handle complicated problems with imprecise inputs. It is suitable for many scientific and real life applications (Mc-Bratney and Moore, 1985). (Bardossy et al., 1995) stated fuzzy as a suitable technique for meteorological prediction or climate classification and described classification of various atmospheric parameters using fuzzy rules. The effect of the fuzzy logic approach based prediction model for temperature, humidity index forecasting was discussed in (Mitra et al., 2006).

(Abdul-Kader ,2009) discussed on application of Multilayer Perceptron (MLP), Radial Basis Function network (RBF) and feed forward neural networks techniques with dissimilar training sets for predictive analysis of Cairo metropolis. (Maqsood et al.,2004) discussed on neural networks based ensemble models for hourly weather forecast of the southern region around Canada using the parameters of temperature, wind speed and relative humidity. Empirical results revealed that RBF network as a suitable weather prediction model. The RBF network performed better than MLP, Elman recurrent neural network, Hop field

model and regression techniques. Kira and Rendell (1992) reported that feature selection is essential to speed up learning. The proposed model consists of parameter selection module and supervised learning (training) module. In the first module, an information gain based parameter ranking is applied for selecting the significant parameters for improving the rainfall prediction efficiency. In the next module the classifiers are trained using selected parameters and complete parameter. The feature selection techniques are effective in modeling daily rainfall prediction (Sudha and Valarmathi, 2014).

This paper is organized as follows: Section 2 discusses the study area of this scheme. In Section 3, we propose the information gain based parameter selection and fuzzy rule based classification for rainfall prediction. In Section 4, we analyze and compare the existing and proposed schemes in terms of accuracy, error rate, RMSE. Section 5 concludes this paper.

Case Study Area

Coimbatore district of Tamil Nadu State in India is selected for the assessment of rainfall prediction. Coimbatore serves as Manchester of South India; it is located in the extreme western region of Tamil Nadu. Coimbatore district's total region covers 746,800 hectares and 43% of the region is bound to agricultural cultivation. The region's climate is classified as sizzling partial dry. The major agricultural crops in this region are cotton, sugarcane, peanut sorghum, maize, rice and pulses. Rainfall received during southwest monsoon is one of the major factors for the groundwater table sauce, but rainfall source is less when compared to winter monsoon. The study region is one of the most important agricultural and industrial area in the country. Fast and uncontrolled industrial development projects have caused climatological changes in past years, hence raised necessity to conduct assessment of factors influencing the current weather prediction.

Materials and methods

Experimental analysis of rainfall record for the Coimbatore region for a period of 27 years from 1984 to 2013). The raw data set is pre-processed for outlier analysis and for removal of missing attribute values. The decision on rainfall occurrence is influenced by eight atmospheric parameters.

P1(Maximum temperature), P2(Minimum temperature), P3(Relative humidity), P4(Relative humidity2), P5(Wind speed), P6(Solar radiation), P7(Sunshine) and P8(Evapotranspiration). This rainfall dataset consists of two class variables. A class variable 'y' of decision parameter P9(Rainfall) means a rainy day else it is a no rain day.

In order to evolve suitable solution to the current challenge, this research focus on applying fuzzy rule based classification approach. The proposed prediction model is trained and validated using reduced feature input determined using information gain measure.

Stastical feature ranking techniques

Information gain measures the quantity of information in bits about the decision class variable and the related class distribution (Dai and Xu, 2013). Entropy measures the expected reduction in vagueness associated with a random feature (Novakovic, 2009). The entropy measure is considered as a measure of unpredictability. Let H (A) be the entropy measure based on the probability density function of a random parameter 'A'. The training set with observed values of 'A' is partitioned on other parameter 'B'. Then entropy measure of a parameter 'A' before partitioning and there exists a relationship between 'A' and 'B'.

Entropy of variable A before observing B is given in equation 1.

$$H(A) = -\sum_{i=1}^{n} P(A_i) \log_2 (P(A_i)) \quad (1)$$

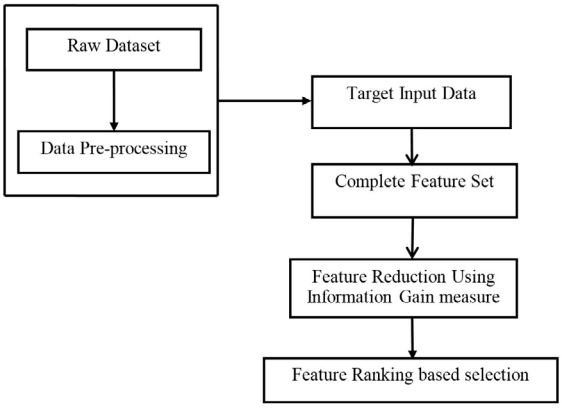
Entropy of variable A after observing B is given in equation 2.

$$H(A_i / B_i) = -\sum_{i-1}^{n} P(B_i) \sum_{i-1}^{n} P(A_i/B_i) \log_2 (P(A_i/B_i)) \quad (2)$$

The information gain measure is estimated as in equation 3.

$$\textit{Information Gain} = H(A_i) - H\left(\frac{A_i}{B_i}\right) \quad (3)$$

The proposed feature selection is an exhaustive task; it requires a suitable stopping criterion to terminate the selection process. The proposed information gain based parameter subset selection module as illustrated in Figure 1. The generation of subset of determined feature reduct is terminated based on problem specific criteria's.



Source: own processing

Figure 1: Information gain based Parameter subset selection module.

The proposed feature selection strategy terminates the subset generation based on the given criteria's:

1. Selection of a predetermined number of features in subset.
2. Achieving a pre-defined number of subsets of the power set.
3. Stopping with respect to evaluation criterion obtained.

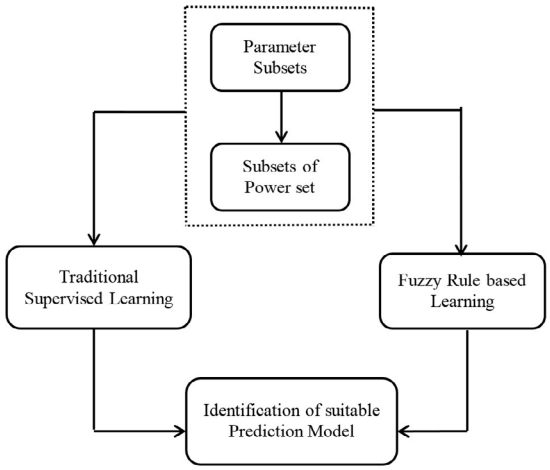
Information Gain	Ranking
0.19326	P4 - RH2
0.12675	P8 - EVP
0.10468	P7 - SS
0.09904	P3 - RH1
0.09814	P2 -MIN
0.08707	P6 -SR
0.06295	P1- MAX
0.00566	P5 - WIND

Source: own processing

Table 1: Information gain based parameter ranking.

The parameters are ranked according to the information gain value from high to low (Table 1). The parameters {P2, P3, P4, P7, P8} having information gain measure equal to or greater than the average information gain value are selected a suitable feature from the complete feature vector. Later, an exhaustive search approach based on the power set algorithm is determined to find possible combinations of subsets of selected significant parameters.

The proposed rule based supervised learning model evaluable process as represented in Figure 2 the performance of set of proposed fuzzy and non-fuzzy rule based classifiers are analyzed in terms of error rate. One of the reasons for using fuzzy logic models is that they can handle vagueness and uncertainty. It deals with arithmetical output and does not require complicated arithmetic enabling fuzzy model as a suitable technique in most of the classification.



Source: own processing

Figure 2: Fuzzy and non-fuzzy supervised learning model evaluation.

It is assumed that in this rainfall classification problem consists of 'w' training model, such that xp = (Xm_1. ….., Xm_n), m = 1, 2. . . W, labeled with one of two possible classes. When RF = 'n' when no rainfall otherwise RF = 'y' is rainfall occur, where Xi is the i^{th} parameter value (i = 1, 2 . . . n) of the training model.

Let R_j be a fuzzy rule represented as:

IF V_1 is P^1_j and P^2_j. . . P^n_j . Van is P^n_j => Cj (class) with (RW_j) (4)

Where Rj is the label of the j^{th} rule, V = (V_1. . . V_n) is a n-dimensional sample vector, A_j is an predecessor fuzzy set, C_j is a class label and RW_j is the rule weight and Fuzzy rules for one particular class with a rule weight (RW_j) associated with this category variable are referred as consequent (Ishibuchi and Yamamoto, 2005).

Fuzzy unordered rule induction algorithm

Fuzzy unordered rule induction algorithm or FURIA is a modification and extension of the ripper rule learner algorithm (Bardossy et al., 1995). FURIA find out to separate each class from all other classes and avoids the default rule set to implement the novel rule stretch approach. Rule stretching is achieved by deleting one or more of its antecedents to generate FURIA's unordered rule set to simplify new queries. It learns

an initial rule set on whole training data and applies pruning for creating new rules for replacement of antecedents without removing all antecedents.

In FURIA, a fuzzy rule obtained by replacing crisp intervals by fuzzy intervals. FURIA implements fuzzy sets with trapezoidal membership function and rules are generated using the greedy approach (Huhn and Hullermeier, 2009). Within Fuzzy rules, traditional crisp boundaries of a rule are substituted by soft boundaries. FURIA represents a fuzzy rule as in equation (6.4).

Fuzzy rules are characterized by its core and its support. Let P be a universal set, with set of instances denoted by *p*, then a fuzzy set F_z in P is a set of ordered pairs in represented as $\{f_z\}$ where, $F_z = \{ (p, \mu_{F(x)}) \mid p$ belong to P$\}$,where $\mu_{F(p)}$ is the membership function of *p* in F which maps *p* to the membership space [0,1].

The grade of membership is assigned 'one' or '0' to those objects that completely belong to Fz and 'zero' or '1' to those that not belong to Fz at all.

The trapezoidal membership function is as below,

$$\mu_F(x, a, b, c, d) = \begin{pmatrix} [0, & \text{if } p < a] \\ [(p - a) / (b - a), & \text{if } a \le p \le b] \\ [1, & \text{if } b < p < c] \\ [(d - p) / (d - c), & \text{if } c \le p \le d] \\ [0, & \text{if } d < p] \end{pmatrix}$$

A fuzzy set P^n_j using an interval of trapezoidal membership function is specified by four parameters (IF = ($\{\emptyset^{S,L}, \emptyset^{S,U}, \emptyset^{C,L}, \emptyset^{C,U}$) Huhn and Hullermeier (2009).

$\{\emptyset^{S,L}, \emptyset^{S,U}\}$ are lower and upper bound of the support elements with membership > 0

$\{\emptyset^{C,L}, \emptyset^{C,U}\}$ are lower and upper bound of the core elements with membership 1.

Let (RH2 ≤ 52) => RF = n be a crisp rule, this rule is valid only when (RH2 ≤ 52) and invalid if (RH2 > 52) for a crisp rule the boundaries are always sharp. Fuzzy rule for the above crisp rule is: (RH2 [-inf, -inf, 52, 53]) => RF= n (CF = 0.91). Implies that the rule is valid if (RH2 ≤ 52), invalid for (RH2 > 53). It is partially valid in between [52 - 53] having soft boundaries.

Crisp rule is defined as a fuzzy rule only if each of its statistical features appears in more than one and a maximum of two predicates in its predecessor part. The FURIA rules are generated using WEKA software.

Relation: {P2, P3, P4, P7, p8}
Instances: 10000
Attributes: 6
Test mode: Ten-fold cross-validation
Classifier model (full training set): FURIA based supervised learning.

FURIA Rules:

- (RH2 in [-inf, -inf, 52, 53]) => RF=n (CF = 0.91)
- (RH2 in [-inf, -inf, 59, 60]) and (EVP in [3.4, 3.5, inf, inf]) and (MIN in [-inf, -inf, 21.5, 21.6]) and (RH1 in [-inf, -inf, 89, 90]) => RF=n (CF = 0.98)
- (RH1 in [-inf, -inf, 92, 93]) and (EVP in [5.5, 5.6, inf, inf]) and (MIN in [23.7, 23.8, inf, inf]) => RF=n (CF = 0.94)
- (EVP in [2.9, 3, inf, inf]) and (RH1 in [-inf, -inf, 92, 93]) and (SS in [4.4, 4.5, inf, inf]) => RF=n (CF = 0.91)
- (EVP in [2.7, 2.8, inf, inf]) and (RH2 in [-inf, -inf, 64, 65]) and (MIN in [-inf, -inf, 20.8, 21]) => RF=n (CF = 0.97)
- (RH2 in [-inf, -inf, 59, 60]) and (RH1 in [-inf, -inf, 92, 93]) and (RH1 in [5, 82, inf, inf]) => RF=n (CF = 0.89)
- (EVP in [2.8, 3, inf, inf]) and (RH1 in [-inf, -inf, 93, 95]) and (SS in [1.2, 1.5, inf, inf]) and (RH2 in [84, 85, inf, inf]) and (RH1 in [88, 89, inf, inf]) and (MIN in [21.4, 21.5, inf, inf]) => RF=n (CF = 0.94)
- (RH2 in [53, 54, inf, inf]) and (RH1 in [92, 93, inf, inf]) => RF=y (CF = 0.73)
- (RH2 in [54, 60, inf, inf]) and (EVP in [-inf, -inf, 1.7, 1.8]) => RF=y (CF = 0.91)
- (RH2 in [67, 68, inf, inf]) and (SS in [-inf, -inf, 4.4, 4.5]) and (RH2 in [-inf, -inf, 83, 84]) => RF=y (CF = 0.76)
- (RH2 in [56, 57, inf, inf]) and (EVP in [-inf, -inf, 6.7, 6.8]) and (MIN in [21.7, 22, inf, inf]) and (RH1 in [-inf, -inf, 82, 83]) and (SS in [-inf, -inf, 3.7, 3.8]) => RF=y (CF = 0.71)
- (RH2 in [45, 47, inf, inf]) and (EVP in [-inf, -inf, 2.9, 3]) and (MIN in [21.3, 21.4, inf, inf]) => RF=y (CF = 0.81)
- (RH2 in [47, 48, inf, inf]) and (MIN in [21.6, 21.7, inf, inf]) and (EVP in [-inf, -inf, 5.1, 5.2]) and (RH1 in [-inf, -inf, 77, 78]) => RF=y (CF = 0.69)

- (RH2 in [45, 48, inf, inf]) and (RH1 in [90, 91, inf, inf]) and (MIN in [22, 22.2, inf, inf]) and (EVP in [-inf, -inf, 3.3, 3.4]) and (EVP in [3, 3.1, inf, inf]) and (MIN in [-inf, -inf, 23.4, 23.5]) => RF=y (CF = 0.89)

Number of Rules: 14

FURIA Rule: (EVP in [2.9, 3, inf, inf]) and (RH1 in [-inf, -inf, 92, 93]) and (SS in [4.4, 4.5, inf, inf]) => RF=n (CF = 0.91).

Let us examine one of the above fuzzy rule generated by FURIA for rainfall prediction, (EVP in [2.7, 2.8, inf, inf]) and (RH2 in [-inf, -inf, 64, 65]) and (MIN (Minimum temperature) in [-inf, -inf, 20.8, 21]) => RF=n (CF = 0.97). The above fuzzy rule is a stretched fuzzy rule from the generated rule set. EVP, RH2 and MIN are attributes of rainfall prediction statistics. The parameter interval range is between 2.7 to 2.8, 64 to 65 and 20.8 to 21. Operator [inf, − inf] points to the interval that has the last valid values. CF indicates the confidence factor of the rule Huhn and Hullermeier (2009).

Results and discussion

The parameter set {P2, P3, P4, P7, P8} and its subsets having three and more of parameters are analysed. The accuracy rate; root means squared error and misclassification rate determine the performance of the classifier. WEKA tool is used for conducting the experimental analysis. It is a good open source machine learning and data mining tool for a broad range of applications Witten and Frank (2005). A detailed experimental study is conducted to evaluate the performance of simple data mining techniques and fuzzy learning algorithms before and after parameter selection. The performance of naive bayes, bayes net, radial basis function Network, sequential minimal optimization and voted perceptron was analysed. For fuzzy based classification, fuzzy rough neural network (FR-NN), fuzzy neural network (F-NN), fuzzy ownership, fuzzy discernibility classifier (F-DC) and Fuzzy unordered rule induction algorithm (FURIA) was evaluated. The accuracy rate of each classifier is determined using confusion matrix. The True positive (T_p), True negative (T_n), fake positive (F_p) and fake negative (F_n) values are represented using this confusion matrix. The accuracy rate is the percent of instances that are correctly classified by the classifier for the specified test set. The incorrectly classified instances determine the error rate or misclassification rate of the classifier. The learning models are evaluated based on the measures in equation (5 to 8).

$$\text{Accuracy Rate } (Ac^R) = \frac{Tp + Tn}{(Tp+Tn+Fp+Fn)} \quad (5)$$

$$\text{Misclassification Rate } (Mc^R) \text{ or Error Rate} = \frac{Fp + Fn}{(Tp+Tn+Fp+Fn)} \quad (6)$$

$$\text{Sensitivity True Positive Rate } (Se^R) = \frac{Tp}{(Tp+ Fn)} \quad (7)$$

$$\text{Specificity True Negative Rate } (Sp^R) = \frac{Tn}{(Tn + Fp)} \quad (8)$$

Performance evaluation pre-parameter subsets selection

The confusion matrix for traditional and fuzzy rule based learning algorithms for the complete parameter is shown in Table 2.

Classifier	TP	FN	FP	TN
FR- NN	6847	897	1026	1230
F- NN	7192	552	1151	1105
FO	6948	796	1077	1179
F- DC	7229	515	1203	1053
FURIA	7311	438	1153	1053
NB	6729	1015	796	1460
BN	6631	1113	675	1581
RBF	7022	722	1044	1212
SMO	7379	365	1319	937
VP	7400	344	1558	698

Source: own processing

Table 2: Confusion matrix before parameter selection.

Among the fuzzy techniques FURIA has acquired high prediction accuracy; the other non-fuzzy models have obtained prediction accuracy almost in analogous range. But when compared with all the models under evaluation FURIA has attained the peak prediction accuracy achieved 83.64%. The accuracy rate, misclassification rate, sensitivity and specificity rate acquired by other classification techniques and FURIA for complete parameter set is shown in Table 3. The selected techniques are trained using the reduced parameter subset obtained using the information gain (the statistical measure). The confusion matrix, accuracy rate, misclassification rate, sensitivity and specificity rate attained by other classification techniques

and FURIA for reduced parameter set is shown in Table 4 and 5.

Classifier	Ac^R(%)	RMSE	Mc^R(%)	Se^R(%)	Sp^R(%)
FR- NN	80.77	0.38	0.19	0.88	0.54
F- NN	82.97	0.41	0.17	0.92	0.48
FO	81.27	0.38	0.18	0.89	0.52
F- DC	82.82	0.32	0.17	0.93	0.46
FURIA	83.64	0.32	0.16	0.94	0.46
NB	81.89	0.36	0.18	0.86	0.64
BN	82.12	0.37	0.17	0.85	0.70
RBF	82.34	0.34	0.17	0.90	0.53
SMO	83.16	0.41	0.16	0.95	0.41
VP	80.98	0.43	0.19	0.95	0.30

Source: own processing

Table 3: Accuracy rate of classifiers before parameter selection.

Performance Evaluation Post - Parameter Subset Selection

Classifier	TP	FN	FP	TN
FR- NN	6743	1001	1072	1184
F- NN	7172	572	1073	1183
FO	6980	764	1060	1196
F- DC	7173	571	1063	1193
FURIA	7309	438	1153	1103
NB	6810	934	844	1412
BN	6779	965	722	1534
RBF	7334	410	1325	931
SMO	7392	352	1389	867
VP	7355	389	1395	861

Source: own processing

Table 4: Confusion matrix after parameter selection.

Classifier	Ac^R (%) before Parameter Selection	Ac^R (%) after Parameter Selection
F- NN	82.97	83.55
FO	81.27	81.76
F- DC	82.82	83.66
FURIA	83.66	84.10
NB	81.89	82.22
BN	82.12	83.13
RBF	82.34	82.65
VP	80.98	82.16

Source: own processing

Table 5: Accuracy rate of Classifiers after Parameter Selection.

Later, an exhaustive subset generation method is implemented to compute the possible subsets of the reduced set obtained using information gain filter. There is no improvement in prediction accuracy of the models when trained with subsets, and few have shown reduced outcomes as indicated in Table 6.

Classifier	Ac^R (%)	RMSE	Mc^R(%)	Se^R (%)	Sp^R(%)
FR- NN	79.27	0.39	0.20	0.87	0.52
F- NN	83.55	0.40	0.16	0.92	0.52
FO	81.76	0.36	0.18	0.90	0.53
F- DC	83.66	0.33	0.16	0.92	0.52
FURIA	84.10	0.37	0.15	0.94	0.49
NB	82.22	0.35	0.17	0.87	0.62
BN	83.13	0.35	0.16	0.87	0.67
RBF	82.65	0.35	0.17	0.94	0.41
SMO	82.59	0.41	0.17	0.95	0.38
VP	82.16	0.42	0.17	0.94	0.38

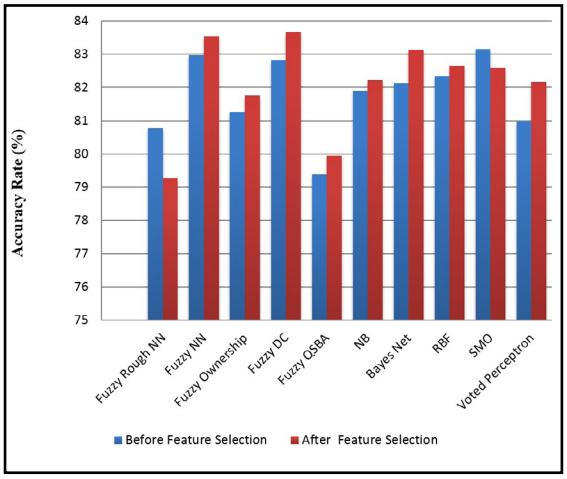
Source: own processing

Table 6: Accuracy rate after parameter selection for subset with 4 parameters.

Classifier	Parameter Subset of Power Set AcR (%)				
	{P-2347}	{P-238}	{P-2378}	{P-2478}	{P-3478}
FR- NN	78.08	79.98	78.7	78.14	77.49
F- NN	82.15	83.55	82.65	82.86	82.69
FO	80.86	82.63	81.81	81.85	81.27
F- DC	82.4	83.80	82.88	83.18	82.91
FURIA	78.67	82.86	79.82	79.33	78.22
NB	81.43	83.02	80.8	81.52	82.11
BN	82.27	83.64	82.81	82.09	82.65
RBF	81.98	82.89	81.91	81.92	82.01
SMO	81.97	82.29	77.44	82.25	82.25
VP	80.88	81.75	80.37	82.46	82.36

Source: own processing

Table 7: Classifiers with enhanced accuracy rate after parameter selection.

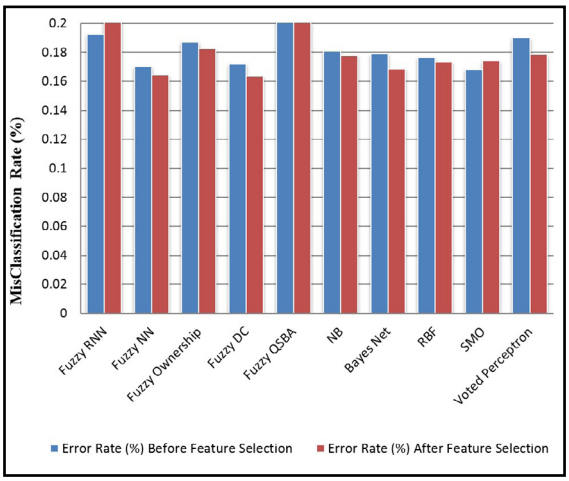


Source: own processing

Figure 3: Visualization of classifiers accuracy rate before and after parameter selection.

From the observed outcomes as shown in Figure 3 and 4, it is concluded that 80% of the learning models have shown better results. Most of the supervised learning algorithms have improved its prediction than before parameter selection. The subsets of power set evaluation notify that there is no substantial improvement in results.

The performance of the model under investigation was not satisfactory when using subsets of the optimal parameter set {P2, P3, P4, P7, P8} identified using information gain. Later, the stopping criterion for subset generation is determined by the learning algorithms accuracy rate. The process of subsets generation process for the complete feature set is terminated based on the accuracy achieved by the subsets.



Source: own processing

Figure 4: Visualization of Classifiers misclassification rate before and after parameter selection.

Conclusion

Feature selection using information gain filter has identified minimum temperature (P2), relative humidity1 (P3), relative humidity2 (P4), the sunshine (P7) and evapotranspiration (P8) as useful parameters for rainfall prediction. Experimental study and evaluations indicate that most of the classification models have shown improved prediction accuracy when trained using feature subset than when trained with a complete feature set as in Table 7. The empirical results have shown that classification algorithms have acquired no significant improvement in accuracy rate when trained with subsets of the reduct set {p2, p3, p4, p7, p8}.

Except Fuzzy Rough Neural network and SMO, other eight classification approaches achieved higher accuracy and lower misclassification rate using parameter selection using information gain measure for feature ranking. FURIA outperformed achieving 84.10% accuracy rate. Therefore fuzzy unordered rule induction is concluded as the suitable classification model for this rainfall prediction statistics.

This detailed analysis and experimental outcomes indicate that supervised learning model results after applying appropriate parameter selection can certainly improve the overall performance of proposed prediction model for optimal reduct set and not for all reduct combinations. More over the prediction accuracy achieved by this fuzzy model is not satisfactory for real time scenarios. Hence forth the necessity of identifying more suitable hybrid intelligent techniques is recommended for modeling the real-time weather prediction system.

Feature selection is a most influencing factor in data mining and decision support systems. Identifying effective inputs for achieving better outcomes will support for effective strategic decisions on various scientific applications. It is proposed to conduct detailed study of other feature selection approaches for determining the effective weather parameters. Apart from statistical measures, feature selection can be achieved using bio inspired procedures. In future we propose to achieve optimal parameter selection using particle Swarm Optimization (PSO), Ant Colony Optimization (ACO) and hybridization of evolutionary and fuzzy approach. The next objective is to propose a fusion of hybrid fuzzy, neural network and evolutionary approach to achieve better prediction accuracy.

Corresponding author:
Dr. M. Sudha
Associate Professor, School of Information Technology and Engineering
Vellore Institute of Technology, India
Phone: +91 9443744781, e-mail msudha@vit.ac.in

References

[1] Abdul-Kader, H. M. (2009) "Neural networks training based on differential evolution algorithm compared with other architectures for weather forecasting", *International Journal of Computer Science and Network Security*, Vol. 9, No.3, pp. 92-99. ISSN 1738-7906.

[2] Al-Matarneh, L., Sheta, A., Bani-Ahmad, S., Alshaer, J. and Al-oqily, I. (2014) "Development of temperature based weather forecasting models using neural networks and fuzzy logic", *International Journal of Multimedia and Ubiquitous Engineering*, Vol. 9, No. 12, pp. 343-366. ISSN 1975-0080. DOI 10.14257/ijmue.2014.9.12.31.

[3] Bardossy, A., Duckstein, L. and Bogardi, I. (1995) "Fuzzy rule based classification of atmospheric circulation patterns", *International Journal of Climatology*, Vol. 15, pp. 1087-1097. ISSN 1097-0088. DOI 10.1002/joc.3370151003.

[4] Blum, A. L. and Rivest, R. L. (1992) Training 3-node neural networks is NP-complete, *Neural Networks,* Vol. 05, pp. 117–127. ISSN 0893-6080.

[5] Dai, J. and Xu, Q. (2013) "Attribute selection based on information gain ratio in fuzzy rough set theory with application to tumour classification", *Applied Soft Computing*, Vol. 13, No. 1, pp. 211-221. ISSN 1568-4946. DOI 10.1016/j.asoc.2012.07.029.

[6] Huhn, J. and Hullermeier, E. (2009) "FURIA: An algorithm for unordered fuzzy rule induction", *Data Mining and Knowledge Discovery*, Vol. 19, No. 3, pp. 293-319. ISSN 1384-5810. DOI 10.1007/s10618-009-0131-8.

[7] Ishibuchi, H. and Yamamoto, T. (2005) "Rule weight specification in fuzzy rule-based classification systems", *IEEE Transaction on Fuzzy Systems*, Vol. 13, No. 4, pp. 428-435. ISSN 1063-6706.

[8] Ishibuchi, H. and Nakashima, T. (2001) "Effect of rule weights in fuzzy rule-based classification systems", *IEEE Transactions on Fuzzy Systems*, Vol. 9, No. 4, pp. 506-515. ISSN 1063-6706.

[9] Kira, K. and Rendell, L. A. (1992) "The feature selection problem: Traditional methods and a new algorithm", *10th National Conference on Artificial Intelligence (AAAI-92)*, San Jose, California, pp. 122-126.

[10] Lee, J., Kim, J., Lee, J. H. I., Cho, I., Lee, J. W., Park, K. H. and Park, K. (2012) "Feature selection for heavy rain prediction using genetic algorithms", *International Symposium on Advanced Intelligent Systems*, Kobe, Japan, pp. 830-833.

[11] Li, K. and Liu, Y. (2005) "A rough set based fuzzy neural network algorithm for weather prediction", *Proceedings of International Conference on Machine Learning and Cybernetics*, Guangzhou, pp. 1888-1892.

[12] Liu, J. N. K., Li, B. N. L. and Dillon, T. S. (2001) "An improved Naive Bayesian classifier technique coupled with a novel input solution method", *IEEE Transactions on Systems, Man and Cybernetics*, Vol. 31, No. 2, pp. 249-256. ISSN 2168-2216. DOI 10.1109/5326.941848.

[13] Maqsood, I., Khan, M. R. and Abraham, A. (2004) "An ensemble of neural networks for weather forecasting", *Neural Computing and Application*, Vol. 13, No. 2, pp. 112-122. ISSN 1433-3058. DOI 10.1007/s00521-004-0413-4.

[14] McBratney, A. and Moore, A. (1985) "Application of fuzzy sets to climatic classification", *Agricultural and Forest Meteorology*, pp. 165-185. ISSN 0168-1923. DOI 10.1016/0168-1923(85)90082-6.

[15] Mitra, A., Meena, L. and Giri, R. (2006) "Forecasting of temperature-humidity index using fuzzy logic approach", *National Conference on Advances in Mechanical Engineering (AIME),* January 2006.

[16] Nikam, V. B. and Meshram, B. B. (2013) "Modeling rainfall prediction using data mining method a bayesian approach", *5th International Conference on Computational Intelligence, Modelling and Simulation*, Seoul, Korea. pp. 132-136.

[17] Niksaz, P. and Latif, A. M. (2014) "Rainfall events evaluation using adaptive neural fuzzy inference system", *International Journal of Information Technology and Computer Science*, Vol. 9, pp. 46-51. E-ISSN 2074-9015, ISSN 2074-9007. DOI 10.5815/ijitcs.2014.09.06.

[18] Novakovic, J. (2009) "Using information gain attribute evaluation to classify sonar targets", *17th Telecommunications forum*, Serbia, Belgrade. pp. 1351-1354.

[19] Seo, J. H., Lee, Y. H. and Kim, Y. H. (2014) "Feature selection for very short-term heavy rainfall prediction using evolutionary computation", *Advances in Meteorology*. Vol. 2014, 15 p. ISSN 1943-5584. DOI 10.1155/2014/203545.

[20] Seo, J. H. and Kim, Y. H. (2012) "Genetic feature selection for very short-term heavy rainfall prediction", *Proceedings of the International Conference on Convergence and Hybrid Information Technology*, Daejeon, Korea. pp. 312-322.

[21] Siedlecki, M. and J. Sklansky (1988) "On automatic feature selection", *International Journal of Pattern Recognition and Artificial Intelligence*, Vol. 2, No. 2, pp. 197–220. ISSN 0218-0014. DOI 10.1142/S0218001488000145.

[22] Sudha, M. and Valarmathi, B. (2013) "Exploration on rough set based feature selection", *International Journal of Applied Engineering Research*, Vol. 8, pp. 1555-1556. ISSN 0973-9769.

[23] Sudha, M. and Valarmathi, B. (2014) "Rainfall forecast analysis using rough set attribute reduction and data mining methods", *Agris on-line Papers in Economics and Informatics*, Vol. 4, No. 4, pp. 145-154. ISSN 1804-1930.

[24] Sudha, M. and B. Valarmathi (2015) "Impact of hybrid intelligent computing in identifying constructive weather parameters for modeling effective rainfall prediction", *Agris on-line Papers in Economics and Informatics*, Vol. 7, No. 4, pp. 151-160. ISSN 1804-1930.

[25] Sudha, M. and B. Valarmathi (2016) "Identification of effective features and classifiers for short term rainfall prediction using rough set based maximum frequency weighted feature reduction technique", *Journal of Computing and Information Technology,* Vol. 24, No. 2, pp. 181-194. ISSN 1846-3908.

[26] Weka Software (2015) [Online]. http://www.cs.waikato.ac.nz/ml/weka/ [Assessed: August 20, 2015].

[27] Witten, I. H. and E. Frank (2005) "*Data Mining: Practical Machine Learning Tools and Techniques*", Morgan Kaufmann, San Francisco, p. 525.

[28] Yu, L. (2005) "Toward integrating feature selection algorithms for classification clustering", *IEEE Transactions on Knowledge and Data Engineering*, Vol. 17, No. 4.

[29] Zadeh, L.A. (1965) "Fuzzy Set", *Information and Control*, Vol. 8, pp. 338-353. DOI 10.1016/S0019-9958(65)90241-X.