The International Journal of Soft Computing and Software Engineering [JSCSE], Vol. 3, No. 3, Special Issue:
The Proceeding of International Conference on Soft Computing and Software Engineering 2013 [SCSE'13],
San Francisco State University, CA, U.S.A., March 2013
Doi: 10.7321/jscse.v3.n3.103 e-ISSN: 2251-7545

An Efficient Decision Tree Classifier to Predict Precipitation Using Gain Ratio

Narasimha Prasad LV
Vardhaman College of Engineering
Hyderabad, India
e-mail: lvnprasad@yahoo.com

Naidu MM
Vignan University, Vadlamudi
Guntur, India
e-mail: mmnaidu@yahoo.com

Abstract—**The population of the world has been increasing substantially. The populous countries like India, seriously lagging behind to provide the basic needs to the people. Food is one of the basic needs that any country has to fulfill. Agriculture is one of the major sectors on which one third of Indian population depends on. The irrigation based countries like India where the water has been the basic resource that forges the plants' growth. The main resource for the irrigation is rainfall which is scientifically a liquid form of precipitation. The atmospheric nimbus clouds are responsible for this precipitation. Prediction of the precipitation is necessary, as it has to be considered during the financial planning of a country. The meteorological departments of every nation are very keen in recording the datasets of precipitation which are huge in content. Hence, data mining is found to be an apt tool which would extract the relation between the datasets and their attributes. A Supervised Learning in Quest is one such data mining algorithm which is eventually a decision tree used to predict the precipitation based on the historical data. The Supervised Learning in Quest decision tree using gain ratio is a statistical analysis for establishing the relation between attribute set and precipitation which furnishes the prediction with an accuracy of 77.78%.**

Keywords- Data Mining; Decision Tree; Meteorology; Precipitation; Prediction; Rainfall; SLIQ;

I. INTRODUCTION

The growth of population is one of the major factors that affect adversely the growth of economy of a country. It is essential to ensure adequacy of infrastructure for providing the basic needs of the growing population. The agricultural sector provides the most of the raw materials required for providing products to meet the basic needs. It is obvious that the agricultural productivity depends on water availability wherein precipitation is the primary source of water. The precipitation is due to the thick layers of the clouds in the atmosphere, which would have attained the melting point [26]. The prediction of the precipitation forms a basis for planning economy with improved accuracy. Hence, there is a need to propose the models for improving accuracy in the precipitation prediction.

A mathematical model is an abstract representation of a real-life problem situation. Many mathematical models which represent the real-life problem situations are complex. Hence, solving such complex models involve in performing a large number of arithmetic and logical operations on related data. The invention of the computer improved accuracy and minimized the time in performing those operations. The prediction of precipitation is a complex and uncertain phenomenon that results in the complex mathematical models. The most of the prediction models employ the huge historical data. Here, data mining can be used for predicting the precipitation more accurately.

Data mining tools can be employed in the fields of prediction constitute artificial neural networks, genetic algorithms, ruled based induction, nearest neighbor method, memory based reasoning, linear discriminate analysis and decision trees. The success rate for the prediction of the precipitation by employing different data mining tools reported in the literature is 43.6% [29]. Recently, Prasad et. al proposed to employ Supervised Learning In Quest (SLIQ) decision tree using Gini index for the prediction of the precipitation which resulted in an accuracy of 72.3% [2]. This paper proposes to employ SLIQ decision tree using gain ratio that improves the accuracy from 72.3% to 77.78%.

The rest of the paper is organized as follows: Section II describes relevant work. Section III provides the information about Decision Trees. In section IV, a brief description about the SLIQ Decision tree algorithm is discussed. Section V describes the rules for decision tree. Section VI describes the experimental results. In section VII conclusions are presented and finally in section VIII, the future enhancements are illustrated.

II. RELEVANT WORK

Research is a continuous process. If anyone imagines that the research on any field is completed and then he/she has to rephrase his/her word of sentence. The research continues beyond this point. In the literature, there are many research findings which are reported for predicting the precipitation with accurate possible rate. Some of them used the traditional methods of the artificial neural networks for the prediction while other methods include the recent developments like Image Processing, Linear Regression and Fuzzy logic and so on.

The International Journal of Soft Computing and Software Engineering [JSCSE], Vol. 3, No. 3, Special Issue:
The Proceeding of International Conference on Soft Computing and Software Engineering 2013 [SCSE'13],
San Francisco State University, CA, U.S.A., March 2013
Doi: 10.7321/jscse.v3.n3.103 e-ISSN: 2251-7545

Frank Silvio Marzano, Giancarlo Rivolta, Erika Coppola, Barbara Tomassetti and Marco Verdecchia used a fully neural network approach to the rainfall field Nowcasting from infrared and microwave passive-sensor imagery aboard [6]. K.Richards and G.D. Sullivan, combined the features of Bayesian scheme for texture analysis of the cloud images which are taken from the ground [7]. C. Jareanpon, W. Pensuwon, R.J. Frank and N. Davey formed radial basis function neural network with a specially designed genetic algorithm [8]. K. Ochiai, H. Suzuki, S. Suzuki, N. Sonehara and Y. Tokunaga stated that the computational time for learning with an acceleration algorithm can be reduced about 10 percent by introducing a pruning algorithm [9]. I.F. Grimes, E. Coppola, M. Verdecchia and G. Visconti presented an approach to cold cloud duration imagery derived from meteosat thermal infrared imagery is used in conjunction with numerical weather model analysis data as an input to an ANN [10]. Thiago N. de Castro, Francisco Souza, Jose M.B. Alves, Ricardo S.T. Pontes, Mosefran B.M. Firmino and Thiago M. de Pereria forecasted seasonal Rainfall using Neo-Fuzzy neuron model [11]. Tuan Zea Tan, Gary Kee Khoon Lee, Shie-Yui Liong, Tian Kuay Lim, Jiawei Chu and Terence Hung IEEE treated the series of rainfall as a continuous time series [12]. Jiansheng Wu Integrated linear regression with ANN. The linear regression extracts linear characteristics of the rainfall [13]. Hui Qi, Ming Zhang and Roderick A. Scofield developed a Multi-Polynomial High Order Neural Network (M-PHONN) [14]. Wint Thida Zaw and Thinn Thu Naing stated that the Multi variables polynomial regression is one of the statistical regression methods used to describe the complex nonlinear input and output relationships [15]. C. Kidd and V. Levizzani stated that the rainfall is spatially and temporally highly variable [16]. Sanjay D. Sawaitul, Prof. K.P. Wagh and Dr. P.N. Chatur used the parameters of the weather like wind direction, wind speed, humidity, rainfall and temperature and so on for the classification and prediction of the future weather by using the back propagation algorithm [17]. Soroosh Sorooshian, Kuo-lin Hsu, Bisher Imam and Yang Hong made global precipitation estimation from satellite image by using artificial neural networks [18]. Kesheng Lu and Lingzhi Wang used a bagging sampling technique is used to generate the training sets for combination model based on support vector machine for the rainfall prediction [19]. Grant W. Petty and Witold F. Krajewski discussed in their research methods based on infrared, visible and passive microwave radiation measurements [20].

III. DECISION TREE

Decision tree is an advanced knowledge discovery process with minimum time complexity and has an ease in the implementation. It establishes relationship between the various datasets by discovering the hidden patterns among the datasets which are huge and complex [3, 4], [26, 27]. As it is known fact that, "*The only way to get more accuracy is to do more research*", which indicates that more and more research has to be done to gain more accurate results. However the research should be carried out by keeping in mind the cost factor. Hence, the scientists have been improving the decision tree algorithms. The use of decision trees have been raised from normal statistical analysis to an effective tool in data mining, text mining, information retrieval and pattern recognition and so on.

The attributes referred in Table I are humidity, temperature, pressure, wind speed and dew point. The amount of water vapor in the air is referred as humidity is invisible in nature. The temperature is the degree of hot or coldness of a body or environment. The temperature is measured in degree centigrade (^{o}C). Atmospheric pressure is the force per unit area exerted against the land surface by the weight of air above the land surface and it is measured in bars. The velocity at which wind is flowing is referred as the wind speed which it is measured in meters per second by an anemometer. Pressure gradient, Rossby waves, jet streams and local weather conditions mainly affect the wind speed which leads to the destruction. Dew point is the temperature at which the air present in the atmosphere can no longer hold all of the water vapor which is mixed with it and some of the water vapor must condense into liquid water.

As it is an established fact that the precipitation generally depends on the various attributes like humidity, temperature, pressure and wind speed and so on. Let us consider a dataset with the similar attributes namely humidity, temperature, pressure, wind Speed and dew point which influence the rainfall and class label as given in Table I. A decision tree is constructed as shown in Fig.1, for the data given in Table 1.

The Table 1 shows 30 days data of humidity, temperature, pressure, wind speed and dew point along with the class label. This is a part of data from Indian Meteorological Department for 15 years.

The decision tree is an inverted tree with root node representing the entire dataset which is partitioned into various branches. The leaves of the branches represent class label as shown in Fig.1.

TABLE I. TRAINING DATASET

Day	Humidity (H)	Temperature (T)	Pressure (P)	Wind Speed (W)	Dew Point (D)	Class
1	97	24	1005	14	21	Rain
2	85	26	1004	16	21	No Rain
3	91	27	1004	14	21	Rain
4	82	27	1006	16	20	Rain
5	81	26	1007	18	19	No Rain
6	95	26	1007	18	20	Rain
7	95	26	1007	16	20	Rain
8	93	26	1008	18	21	Rain
9	87	24	1005	13	21	Rain
10	88	24	1005	11	21	Rain
11	80	26	1005	14	21	Rain
12	89	26	1005	14	21	Rain

The International Journal of Soft Computing and Software Engineering [JSCSE], Vol. 3, No. 3, Special Issue:
The Proceeding of International Conference on Soft Computing and Software Engineering 2013 [SCSE'13],
San Francisco State University, CA, U.S.A., March 2013
Doi: 10.7321/jscse.v3.n3.103 e-ISSN: 2251-7545

13	86	27	1006	14	21	No Rain
14	86	28	1007	10	22	Rain
15	94	27	1006	14	21	Rain
16	88	26	1004	13	21	No Rain
17	92	27	1005	13	21	Rain
18	86	27	1007	11	21	Rain
19	82	27	1006	11	21	Rain
20	76	27	1007	14	19	No Rain
21	79	27	1008	11	20	No Rain
22	75	27	1008	13	20	No Rain
23	84	27	1007	13	20	No Rain
24	88	26	1006	11	21	Rain
25	86	25	1005	16	19	Rain
26	78	28	1006	13	21	No Rain
27	79	27	1008	13	19	No Rain
28	80	28	1008	8	20	No Rain
29	84	29	1009	6	21	No Rain
30	76	27	1009	6	22	Rain

TABLE II NOTATIONS USED IN PRESENTING SLIQ ALOGRITHM

Symbols	Description
D	Set of training tuples with associated class labels
D_J	The set of data tuples in D satisfying outcome J
|D|	The number training tuples in D
C	The class label
Entropy (D)	The information needed to classify a tuple in D
Splitinfo (V)	Normalization to information gain.
Split point	Midpoint of V_i and V_{i+1}
V	An attribute list
V_i	Set of values in attribute V
V_{i+1}	Changed Class value in attribute V
P_i	The probability that a tuple in D belongs to class C_i
D_i	Values which are greater than or equal to the Split point
D_j	Values which are less than the Split point

The criterion for partitioning dataset at a level is explained in the next section. Decision trees can be used for dataset whether it is continuous or discontinuous. The category of dataset is taken into account which is called as the class label. One of the attributes becomes the root node for the decision tree whereas class label is the leaf node as shown in Fig.1. The knowledge based mining is not so effective in establishing temporal attribute relationships.

IV. SLIQ DECISION TREE ALGORITHM

The decision tree classifier, SLIQ [1] can handle numeric as well as categorical attributes. It employs a pre-sorting technique for reducing the cost of evaluating numeric attributes during the tree-growth phase. Further, the SLIQ using the Minimum Description Length (MDL) principle employs a tree pruning algorithm. It is reported that the SLIQ algorithm is inexpensive in resulting compact and accurate trees [1]. The SLIQ ensures scalability in classifying large datasets consisting of a large number of classes and attributes.

In the construction of the decision tree gain ratio is evaluated at every successive midpoint of the attribute values. However, the efficiency of the SLIQ decision tree algorithm can be improved by evaluating gain ratio only at the midpoints of the attributes where the class information changes. The algorithm for the construction of SLIQ decision tree for the prediction of precipitation is presented below. The notations used are given in Table II.

Overview of SLIQ Decision tree growth and split points

1. Read dataset into the root node of the SLIQ decision tree
2. Generate an attribute list for each attribute of the dataset
3. Sort the attribute lists on attribute value in non-decreasing order
4. Compute the entropy for the root node

$$Entropy(D) = -\sum_{i=1}^{N} P_i \log_2 P_i \quad (1)$$

5. Compute the Info of attribute list 'V'

$$Info(V) = \sum_{j=1}^{N} P_j \left[-\sum_{i=1}^{N} P_i \log_2 P_i \right] \quad (2)$$

6. Compute the Gain for each attribute list

$$Gain\ (V) = Entropy\ (D) - Info\ (V) \quad (3)$$

7. Compute split information for a set of values of attribute 'D_i' and 'D_j'

$$Splitinfo\ (V) = -\left[\frac{|D_i|}{|V|} \log_2 \left(\frac{|D_i|}{|V|} \right) + \frac{|D_J|}{|V|} \log_2 \left(\frac{|D_J|}{|V|} \right) \right] \quad (4)$$

8. Determine the Gain Ratio for the attribute values in attribute list 'V'

$$Gain\ Ratio\ (V) = Gain\ (V) / Splitinfo(V) \quad (5)$$

9. Determine maximum gain ratio from among the gain ratios which become the basis for the best split as shown in Table III.

$$Best\ Split = Max.\ Gain\ Ratio\ value\ of\ attribute \quad (6)$$

10. Partition the root node into leaf nodes based on the best split point
11. Repeat the steps 5 through 10 reading the root node as leaf node until all leaf nodes contain the same class labels.

The primary metric for evaluating the prediction of precipitation is accuracy - the accuracy of a predictor refers to how well a given prediction can give the value of the predicted attribute for new or previously unseen data.

$$Accuracy = Correct\ predictions / Total\ predictions \quad (7)$$

The ideal goal is to produce compact, accurate trees in a short time with scalability - the SLIQ decision tree algorithm used for the prediction of precipitation takes N input attributes with its associated class labels as an input and produces the decision tree along with the rules.

The simulated tree shown in Fig. 1 consists of 13 nodes and 7 out of them are depicting rain and the remaining 6 are depicting no rain. The decision tree shown in Fig. 1 NR indicates no rain and R indicate rain.

The International Journal of Soft Computing and Software Engineering [JSCSE], Vol. 3, No. 3, Special Issue:
The Proceeding of International Conference on Soft Computing and Software Engineering 2013 [SCSE'13],
San Francisco State University, CA, U.S.A., March 2013
Doi: 10.7321/jscse.v3.n3.103 e-ISSN: 2251-7545

FIG. 1. GAIN RATIO BASED DECISION TREE

TABLE III. GAIN RATIO BASED SPLIT VALUE FOR VARIOUS ATTRIBUTES

Iteration	Humidity		Temperature		Pressure		Wind Speed		Dew Point	
	Split Value	Gain Ratio	Split Value	Gain Ratio	Split Value	Gain Ratio	Split Value	Gain Ratio	Split Value	Gain Ratio
Step 1	**86.0**	**0.2791**	28.5	0.2149	1007.5	0.1146	9.0	0.0495	20.0	0.1029
Step 2	83.0	0.1602	27.5	0.1602	**1007.0**	**0.2050**	17.0	0.0976	20.0	0.1602
Step 3	83.0	0.4459	**27.5**	**0.4459**	1006.0	0.205	13.0	0.2367	20.5	0.2367
Step 4	**83.0**	**1.00**	27.0	0.3112	1006.0	0.3112	15.0	0.3112	20.5	0.1510
Step 5	77.0	0.2147	27.5	0.0563	1008.5	0.3677	7.0	0.3677	**20.5**	**0.3677**
Step 6	83.0	-1.0	**27.5**	**1.0**	1007.5	1.0	15.0	1.0	20.5	-1.0
Step 7	88.0	0.0176	25.5	0.0690	**1007.0**	**0.0817**	15.0	00690	20.5	0.0579
Step 8	88.0	0.0452	**25.5**	**0.1425**	1006.0	0.0452	13.0	0.0859	20.5	0.0631
Step 9	**88.0**	**0.5171**	27.0	0.0060	1006.0	0.0060	13.0	0.1284	20.5	-1.0
Step 10	83.0	-1.0	27.0	0.1980	1006.0	0.1188	**13.5**	**0.1908**	20.5	-1.0
Step 11	83.0	-1.0	27.0	0.2740	1006.0	0.2740	**13.0**	**0.2740**	20.5	-1.0
Step 12	83.0	-1.0	**27.0**	**1.0**	1007.5	-1.0	15.0	-1.0	20.5	-1.0

V. RULES FOR DECISION TREE

Once the decision tree is constructed, there is a possibility that the tree is very large to understand. Hence, to simplify the understanding of the large decision tree the rules are generated.

Rule 1: If [(humidity < 86.0) and (pressure < 1007.0) and (temperature < 27.5) and (humidity < 83.0)] Then (Prediction = Rain)

Rule 2: If [(humidity < 86.0) and (pressure < 1007.0) and (temperature < 27.5) and (humidity >= 83.0)] Then (Prediction = NoRain)

Rule 3: If [(humidity < 86.0) and (pressure < 1007.0) and (temperature >= 27.5)] Then (Prediction = NoRain)

Rule 4: If [(humidity <86.0) and (pressure >=1007.0) and (dew-point <20.5)] Then (Prediction=NoRain)

Rule 5: If [(humidity < 86.0) and (pressure >= 1007.0) and (dew-point >= 20.5) and (temperature < 27.5)] Then (Prediction = Rain)

Rule 6: If [(humidity < 86.0) and (pressure >= 1007.0) and (dew-point >= 20.5) and (temperature >= 27.5)] Then (Prediction = NoRain)

Rule 7: If [(humidity >= 86.0) and (pressure < 1007.0) and (temperature < 25.5)] Then (Prediction = Rain)

Rule 8: If [(humidity >= 86.0) and (pressure < 1007.0) and (temperature >= 25.5) and (humidity < 88.0)] Then (Prediction = NoRain)

Rule 9: If [(humidity >= 86.0) and (pressure < 1007.0) and (temperature >= 25.5) and (humidity >= 88.0) and (wind-speed < 13.5) and (wind-speed < 13.0)] Then (Prediction = Rain)

Rule 10: If [(humidity >= 86.0) and (pressure < 1007.0) and (temperature >= 25.5) and (humidity >= 88.0) and (wind-speed < 13.5) and (wind-speed >= 13.0) and (temperature < 27.0)] Then (Prediction = NoRain)

Rule 11: If [(humidity >= 86.0) and (pressure < 1007.0) and (temperature >= 25.5) and (humidity >= 88.0) and (wind-speed < 13.5) and (wind-speed >= 13.0) and (temperature >= 27.0)] Then (Prediction = Rain)

Rule 12: If [(humidity >= 86.0) and (pressure < 1007.0) and (temperature >= 25.5) and (humidity >= 88.0) and (wind-speed >= 13.5)] Then (Prediction=Rain)

Rule 13: If [(humidity >= 86.0) and (pressure >= 1007.0)] Then (Prediction = Rain)

The International Journal of Soft Computing and Software Engineering [JSCSE], Vol. 3, No. 3, Special Issue:
The Proceeding of International Conference on Soft Computing and Software Engineering 2013 [SCSE'13],
San Francisco State University, CA, U.S.A., March 2013
Doi: 10.7321/jscse.v3.n3.103 e-ISSN: 2251-7545

VI. EXPERIMENTAL RESULTS

The data taken for the training needs to be sorted during the initial stage of the tree growth phase of decision tree construction [3]. As per the training data, humidity is the first attribute. Take the humidity attribute and its corresponding class label as a pair, identify the split points whenever there is a change in the class label. The better split point needs to be found for increasing the accuracy of prediction. For every split point identified find the midpoint for the changed class labels and proceed until it reaches the end of the data as shown in Table IV.

From the Table IV it is clearly visible that there is a change in the class label for the first time at the 3rd position. Mark it as split point and take the midpoint value of 2nd and 3rd class label values i.e. midpoint (76, 76) =76. Similarly the second split point occurs at 4th position. Mark it as split point and take the midpoint value of 3rd and 4th class label values i.e. midpoint (76, 78) = 77. Proceeding in this order there are nine split points as the class label is changing at nine positions.

Repeat the procedure to find out the split points for the attribute temperature shown in Table V, attribute pressure shown in Table VI, attribute wind speed shown in Table VII and attribute dew point shown in Table VIII.

TABLE IV. DATASET SORTING ON HUMIDITY

Humidity	Class	Split Point
75	No Rain	
76	No Rain	→ 76
76	Rain	→ 77
78	No Rain	
79	No Rain	
79	No Rain	
80	No Rain	
80	Rain	→ 80.0
81	No Rain	→ 80.5
82	Rain	→ 81.5
82	Rain	→ 83.0
84	No Rain	
84	No Rain	
85	No Rain	
86	No Rain	
86	Rain	→ 86.0
86	Rain	
86	Rain	
87	Rain	→ 87.5
88	No Rain	
88	Rain	→ 88.0
88	Rain	
89	Rain	
91	Rain	
92	Rain	
93	Rain	
94	Rain	
95	Rain	
95	Rain	
97	Rain	

TABLE V. DATASET SORTING ON TEMPERATURE

Temperature	Class	Split Point
24	Rain	
24	Rain	
24	Rain	
25	Rain	→ 25.5
26	No Rain	
26	No Rain	
26	No Rain	→ 26
26	Rain	
26	Rain	
26	Rain	
26	Rain	
26	Rain	
26	Rain	
26	Rain	→ 26.5
27	No Rain	
27	No Rain	
27	No Rain	
27	No Rain	
27	No Rain	
27	No Rain	
27	Rain	→ 27
27	Rain	
27	Rain	
27	Rain	
27	Rain	
27	Rain	
27	Rain	→ 27.5
28	No Rain	
28	No Rain	→ 28.0
28	Rain	→ 28.5
29	No Rain	

TABLE VI. DATASET SORTING ON PRESSURE

Pressure	Class	Split Point
1004	No Rain	
1004	No Rain	
1004	Rain	→ 1004
1005	Rain	
1005	Rain	
1005	Rain	
1005	Rain	
1005	Rain	
1005	Rain	
1005	Rain	→ 1005.5
1006	No Rain	
1006	No Rain	→ 1006
1006	Rain	
1006	Rain	
1006	Rain	
1006	Rain	→ 1006.5
1007	No Rain	
1007	No Rain	
1007	No Rain	
1007	Rain	→ 1007
1007	Rain	
1007	Rain	
1007	Rain	→ 1007.5
1008	No Rain	
1008	No Rain	
1008	No Rain	
1008	No Rain	→ 1008
1008	Rain	→ 1008.5
1009	No Rain	→ 1009
1009	Rain	

The International Journal of Soft Computing and Software Engineering [JSCSE], Vol. 3, No. 3, Special Issue:
The Proceeding of International Conference on Soft Computing and Software Engineering 2013 [SCSE'13],
San Francisco State University, CA, U.S.A., March 2013
Doi: 10.7321/jscse.v3.n3.103 e-ISSN: 2251-7545

TABLE VII. DATASET SORTING ON WIND SPEED

Wind Speed	Class	Split Point
6	No Rain	6
6	Rain	7
8	No Rain	9
10	Rain	10.5
11	No Rain	11
11	Rain	
11	Rain	
11	Rain	
11	Rain	12
13	No Rain	
13	No Rain	
13	No Rain	
13	No Rain	
13	No Rain	13
13	Rain	
13	Rain	13.5
14	No Rain	
14	No Rain	14
14	Rain	
14	Rain	
14	Rain	
14	Rain	
14	Rain	15
16	No Rain	16
16	Rain	
16	Rain	
16	Rain	17
18	No Rain	18
18	Rain	
18	Rain	

TABLE VIII. DATASET SORTING ON DEW POINT

Dew Point	Class	Split Point
19	No Rain	
19	No Rain	
19	No Rain	19
19	Rain	19.5
20	No Rain	
20	No Rain	
20	No Rain	
20	No Rain	20
20	Rain	
20	Rain	
20	Rain	20.5
21	No Rain	
21	No Rain	
21	No Rain	
21	No Rain	
21	No Rain	
21	Rain	21
21	Rain	
21	Rain	
21	Rain	
21	Rain	
21	Rain	
21	Rain	
21	Rain	
21	Rain	
21	Rain	
21	Rain	
22	Rain	

Now, compare all the split points' gain ratio values and the value which is maximum is the best split point for that attribute as shown in Table III. The gain value obtained for the attribute is to be divided by split info value of class label, in order to obtain the gain ratio value for that attribute and is shown in equation (9).

$$Gain\ Ratio\ (V) = Gain\ (V) / Split\ info\ (V) \quad (9)$$

Repeat the above procedure by taking the temperature attribute along with the class label, Pressure attribute along with the class label, wind speed attribute along with the class label and finally dew point attribute along with the class label to get the best split points. Choose the maximum gain ratio value and that itself becomes the root node. Based on the threshold value of the root node generates the tree. Repeat the procedure till it is terminated with a unique class label.

The gain ratio is generally used to measure the inequalities among the statistical data and its frequencies. So far, its use is limited for the analysis of wealth and income of the economic countries. Due to the inequalities present in the probabilities, there may be some error. But, irrespective of its limitation present, it has a wide variety of the applications in statistical analysis.

The gain ratio is used here for the construction of the decision tree where the roots and sub-roots are classified. The use of the gain ratio for the rainfall analysis is quite apt because of the irregularities present in the statistical data of the precipitation. The precipitation data is used does not follow an order in other words a sequential path. This may be due to the inequalities of the present attribute with former attribute. This may change to a great extent or to some extent depending on the Mother Nature.

Some experiments have been conducted on real data to analyze the accuracy of the tree. We have used the dataset from the accuweather.com of Indian Meteorological Department. The goal is to predict the precipitation for rainfall. The dataset consists of 15 years of data from the year 1998 to 2012 containing of 5230 examples. Out of 15 years data 9 years data is used as training dataset and the remaining 6 years data is used as test dataset.

It has been found in Table IX, the distinction between the success rate of prediction and time. It can also be observed, that the maximum efficiency obtained is 74.1% on one year dataset, 77.47% for two years dataset, 77.38% for three years dataset, 77.17% for four years dataset, 77.39% for a five 5 years dataset and 77.78 % for a six years dataset. The average efficiency has been found to be 77.78%. Though, this contributes a decent efficiency or success rate, the other methods of back propagation neural networks [7,8], [12-15], linear discriminate statistical analysis [16] and J48

The International Journal of Soft Computing and Software Engineering [JSCSE], Vol. 3, No. 3, Special Issue:
The Proceeding of International Conference on Soft Computing and Software Engineering 2013 [SCSE'13],
San Francisco State University, CA, U.S.A., March 2013
Doi: 10.7321/jscse.v3.n3.103 e-ISSN: 2251-7545

are analyzed to select the best performing method of prediction of the precipitation.

The published results for this dataset are: 64.3% accuracy for backpropagation, 58% for a linear discriminant and 68.6% for J48. Using the same training and test datasets, Since the average accuracy using SLIQ with gain ratio is 77.78% as shown in Fig. 4, SLIQ using gain ratio can be considered as the best performing method for the prediction of precipitation.

TABLE IX. RESULT SHOWING THE ACCURACY AND TIME OF RESPONSE

No. of records	Correct predictions	In correct predictions	Accuracy (%)	Time (Sec)
363	269	94	74.104	37
728	566	162	77.471	40
995	770	225	76.381	42
1262	974	288	77.175	44
1619	1253	366	77.392	46
1981	1541	440	77.778	47

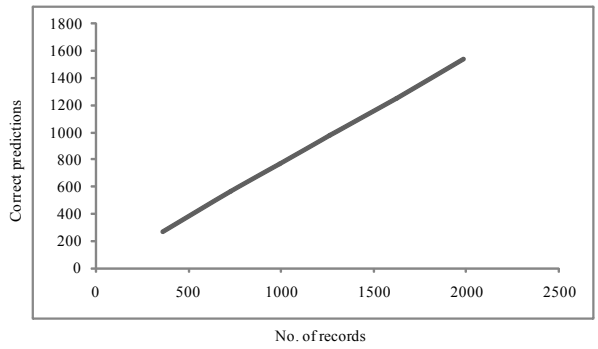

FIG 2. NO. OF RECORDS VS CORRECT PREDICTIONS

FIG 3. NO. OF RECORDS VS INCORRECT PREDICTIONS

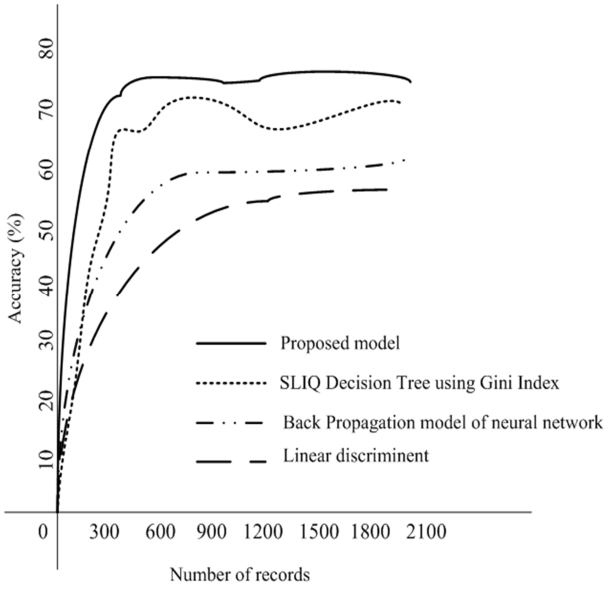

FIG 4. NO. OF RECORDS VS ACCURACY (%)

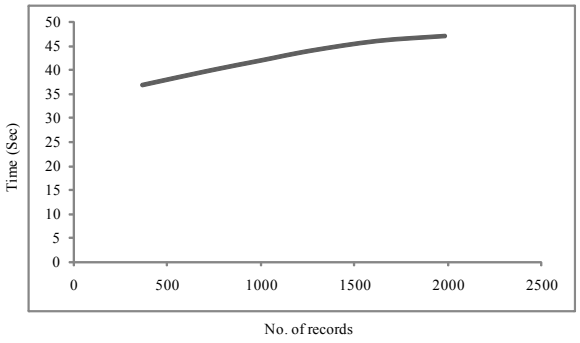

FIG 5. NO. OF RECORDS VS TIME (SEC)

The variation of correct predictions with dataset is shown in Fig. 2. This indicates that there lies a linear relationship between correct predictions and number of records in the dataset.

The variation of incorrect predictions with dataset is shown in Fig.3. This indicates that there lies a non linear relationship between incorrect predictions and number of records in the dataset. From the above graph, the number of incorrect predictions follows a decreasing trend up to 600 records and thereafter increases non linearly. The variation of accuracy with the dataset is plotted in Fig. 4. The variation of time of response towards dataset is plotted in Fig. 5.

VII. CONCLUSION

The economy of a nation depends on agricultural productivity which is the basis for formulating economic policy. The agricultural productivity depends on the availability of water. The precipitation is the major source of water which depends on various attributes like humidity, pressure, temperature, wind speed, dew point and so on.

The International Journal of Soft Computing and Software Engineering [JSCSE], Vol. 3, No. 3, Special Issue:
The Proceeding of International Conference on Soft Computing and Software Engineering 2013 [SCSE'13],
San Francisco State University, CA, U.S.A., March 2013
Doi: 10.7321/jscse.v3.n3.103 e-ISSN: 2251-7545

Hence, the prediction of precipitation becomes a difficult task as it has to consider many parameters. Many techniques such as neural networks, artificial intelligence, used for prediction of precipitation have less accuracy. So far, the maximum accuracy reported is 72.3%. This study employed SLIQ decision tree using gain ratio as splitting criterion. For evaluating the effectiveness of this model the historical data obtained from IMD is applied. It is found that the method proposed in this paper gives higher accuracy when compared to the other models.

VIII. FUTURE ENHANCEMENTS

In this paper, we highlighted gain ratio based SLIQ decision tree algorithm, which gives maximum accuracy. For future implementation various other decision tree algorithms like CART, SPRINT, ELEGANT, EC4.5 with additional parameters can be developed. A decision tree must be developed for the dynamic mode of data rather than static mode.

REFERENCES

[1] Manish Mehta, Rakesh Agrawal, Jorma Rissanen, "SLIQ, A Fast Scalable Classifier for Data Mining," EDBT '96 Proceedings of the 5th International Conference on Extending Database Technology: Advances in Database Technology, Springer-Verlag London, UK, pp.18-32, 1996.

[2] Narasimha Prasad, Prudhvi Kumar Reddy, Naidu MM, "An Approach to Prediction of Precipitation Using Gini Index in SLIQ Decision Tree", 4th International Conference on Intelligent Systems, Modeling & Simulation, Bangkok, pp.56-60, 2013.

[3] Yu-Shan Shih, "Families of Splitting Criteria for Classification Trees", Statistics and Computing, Vol. 9, pp.309-315, 1999.

[4] Mahesh V. Joshi, Eui-Hong (Sam) Han, George Karypis, Vipin Kumar, Parallel Algorithms in Data Mining. CRPC Parallel Computing Handbook, Morgan Kaufmann, 2000.

[5] B. Chandra, P. Paul Varghese, Fuzzy SLIQ Decision Tree Algorithm. IEEE Transactions on Systems. Man and Cybernetics – Part B: Cybernetics, Vol.38, pp. 1294-1301, 2008.

[6] K Frank Silvio Marzano, Giancarlo Rivolta, Erika Coppola, Barbara Tomassetti and Marco Verdecchia, "Rainfall Nowcasting from Multisatellite Passive-Sensor Images Using a Recurrent Neural Network", IEEE Transactions on Geosciences and Remote Sensing, Vol. 45, pp. 3800-3812, 2007.

[7] K.Richards and G.D. Sullivan, "Estimation of cloud cover using colour and texture", British Machine Vision Conference, Springer, pp. 436-442, 1992.

[8] C. Jareanpon, W. Pensuwon, R.J. Frank and N. Davey, "An adaptive RBF network optimized using a Genetic algorithm applied to Rainfall Forecasting", International Symposium on Communications and Information Technologies, pp. 1005-1010, 2004.

[9] K. Ochiai, H. Suzuki, S. Suzuki, N. Sonehara and Y. Tokunaga, "Snowfall and rainfall forecasting from the images of weather radar with artificial neural networks", IEEE International Conference on Neural Networks, pp. 1942-1948, 1995.

[10] D.I.F. Grimes, E. Coppola, M. Verdecchia and G. Visconti, "A Neural Network Approach to Real-time Rainfall Estimation for Africa and using Satellite Data", American Meteorological Society, pp. 1119-1133, 2003.

[11] Thiago N. de Castro, Francisco Souza, Jose M.B. Alves, Ricardo S.T. Pontes, Mosefran B.M. Firmino and Thiago M. de Pereria, "Seasonal Rainfall Forecast using a Neo- Fuzzy Neuron Model", 9th IEEE International Conference on Industrial Informatics, pp. 694-698, 2011.

[12] Tuan Zea Tan, Gary Kee Khoon Lee, Shie-Yui Liong, Tian Kuay Lim, Jiawei Chu and Terence Hung, "Rainfall Intensity Prediction by a Spatial-Temporal Ensemble," IEEE International Joint Conference on Neural Networks, pp. *1721-1727* , 2008.

[13] Jiansheng Wu in "A Novel Nonlinear Ensemble Rainfall Forecasting Model incorporating Linear and Nonlinear Regression", Fourth International Conference on Natural Computation, pp. 18-20, 2008.

[14] Hui Qi, Ming Zhang and Roderick A. Scofield, "Rainfall Estimation using M-PHONN Model", IEEE International Conference on Neural Networks, pp. 1620-1624, 2001.

[15] Wint Thida Zaw and Thinn Thu Naing, "Modeling of Rainfall prediction over Myanmar using Polynomial Regression", International Conference on Computer Engineering and Technology, pp. 11-15, 2009.

[16] C. Kidd and V. Levizzani, "Status of Satellite Precipitation Retrievals", Hydrology and Earth System Sciences, pp. 1109-1116, 2011.

[17] Sanjay D. Sawaitul, Prof. K.P. Wagh, Dr. P.N. Chatur, "Classification and Prediction of Future Weather by using Back Propagation Algorithm - An Approach" International Journal of Emerging Technology and Advanced Engineering, pp. 110-113, 2012.

[18] Soroosh Sorooshian, Kuo- lin Hsu, Bisher Imam and Yang Hong, "Global Precipitation Estimation from Satellite Image using Artificial Neural Networks", Cambridge University Press, pp. 21-28, 2007.

[19] Kesheng Lu and Lingzhi Wang, "A Novel Nonlinear Combination Model Based on Support Vector Machine for Rainfall Prediction" Fourth International Joint Conference on Computational Sciences and Optimization, pp. 1343-1346, 2011.

[20] Grant W. Petty and Witold F. Krajewski, "Satellite Estimation of Precipitation over land", Hydrological Sciences Journal, pp. 433-453, 1996.

[21] K. Richards and G.D. Sullivan, "Estimation of Cloud Cover using Colour and Texture Intelligent Systems Group", University of Reading, RG6 2AY, pp. 436-442, 2006.

[22] Koizumi, K., "An objective method to modify numerical model forecasts with newly given weather data using an artificial neural network, Weather Forecast", 14, pp. 109-118, 1999.

[23] Luk K. C., Ball J. E, Sharma A, A Study of Optimal Model Lag and Spatial Inputs to Artificial Neural Network for Rainfall Forecasting, Journal of Hydrology, vol.227, pp.56-65, 2000.

[24] Robert A. Houze, Cloud Dynamics. Academic Press, 1993.

[25] J.R. Quinlan, Introduction of Decision Tree, Machine Learning, Vol. 1, pp.81-106, 1986.

[26] Wei-Yin Loh, Classification and Regression Tree Methods. Ruggeri, Kenett and Faltin. Wiley, pp.315–323, 2008.

[27] Wang Yong, XU Hong, Guo Zengzhang, Ding Keliang, Liu Yanping, Wen Debao, the Study of Rainfall Forecast Based on Neural Network and GPS Precipitable Water Vapor. IEEE International Conference on Environmental Science and Information Application Technology: pp.17–20, 2010.

[28] Jehangir Ashraf Awan, Onaiza Maqbook, Application of Artificial Neural Networks for Monsoon Rainfall Prediction. IEEE International Conference on Emerging Technologies, pp.27–32, 2010.

[29] Jean Claude Berges, Neural Networks and Tree Classifiers. IEEE International Symposium on Geoscience and Remote Sensing, pp.887–889, 2003.

[30] Yuhui Wang, Yunzhong Jiang, Xiaohui Lei, Wang Hao, Rainfall-Runoff Simulation Using simulated Annealing Wavelet BP Neural Networks, IEEE International Conference on Intelligent Computation Technology and Automation, pp.963–967, 2010.

Doi: 10.7321/jscse.v3.n3.103 e-ISSN: 2251-7545

[31] Jiansheng Wu, A Novel Nonlinear Eensemble Rainfall Forecasting Model Incorporating Linear and Nonlinear Regression. IEEE International Conference on Natural Computation: pp.34–38, 2008.