IJCSE International Journal of Computer Sciences and Engineering Open Access
Research Paper Vol.-6, Issue-10 Oct 2018 E-ISSN: 2347-2693

# Analyzing Machine learning Algorithm for Predicting an Accuracy of Meteorological Data

**M. Manikandan[1*], R. Mala[2]**

[1]Dept. of Computer Science, Marudupandiyar College, Thanjavur, Tamil Nadu, India
[2]Dept. of Computer Science, Alagappa University College, Paramakudi, Tamil Nadu, India

*Corresponding Author: rmgvm2007@yahoo.com, Cell: 9488424357*

**Available online at: www.ijcseonline.org**

Accepted: 25/Oct/2018, Published: 31/Oct/2018

***Abstract***— Meteorological data analysis in the form of data mining is concerned to predict the knowledge of weather condition. To make an accurate prediction is one of the challenging of meteorologist to survey the weather condition efficiently. Decision tree algorithms are suitable for analyzing the data of meteorological behavior. By evaluates three algorithm of decision tree such as Random Forest, C4.5, C4.5 with Bootstrap aggregation, to analyse the time efficiency and accuracy of classification. These accuracy of algorithm when it operates on trained weather data of selected location. Those locations are selected through monsoon condition based on India country.

***Keywords***—Randon Forest, C4.5, C4.5 with Bootstrap Algorithm, Meterological Data, Accuracy,Time efficiency

## I. INTRODUCTION

The research has been focused on retrieval of relevant information from the collection of data is a challenging task in data mining techniques [2,3,4,5,6]. The performance can be identified by the factor of each seasonal climate stage for analyzing in monsoon and occurring weather condition during the season. Each path from the root of a decision tree to one of its leaves can be transformed into a rule simply by conjoining the tests along the path to form the antecedent part, and taking the leaf's class prediction as the class values. Pre-pruning involves deciding when to stop developing sub-trees during the tree building process. The minimum number of observations in a leaf can determine the size of the tree. After a tree is constructed, the C4.5 rule induction program can be used to produce a set of equivalent rules. Pruning produces fewer, more easily interpreted results.

## II. RESEARCH METHODOLOGY

In this research, to classify the quantitative data represents temperature, Vapor pressure, cloud cover and relative humidity and cyclone form for analyzing the weather prediction during the season of summer, winter, northeast, southwest monsoon. Figure.1 illustrates the research methodology.

The relative humidity can be calculated by Vapor pressure and saturation of vapor pressure which is measured by percentage. The data sets are collected from the India Meteorological Department section websites, here the domain variable training data can be represent in Table 1.

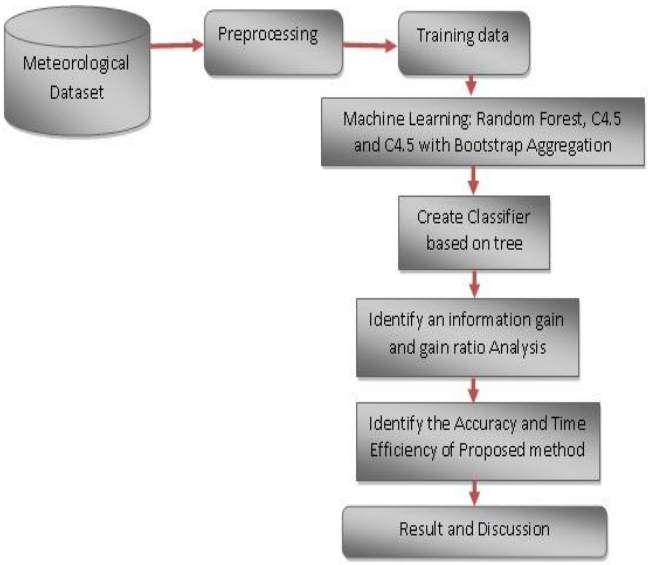


Figure.1 Frame work of Research Methodology

The data sets are collected from the India Meteorological Department section websites, here the domain variable training data can be represent in Table 1. The month can be split into winter seasons are comes January and February, the summer seasons are comes from March to May Southwest periods follows June, July, August, and September and a Northeast monsoon period follows on October November, December.

**Preprocessing -** In Preprocessing, the raw data of numerical data is converted into nominal data for classifying process.

 

After the preprocess, apply three algorithm of machine learning technique of Random forest, C4.5 algorithm and C4.5 Bootstrap aggregation algorithm for analysing the accuracy of classification process.

Table 1 Domain variable of collecting data

| Domain Variables | Abbreviation |
|---|---|
| W_Temp | Temperature of winter seasons in January and February Month |
| W_Cloud cover | Cloud cover of winter seasons |
| W_Vapor pressure | Vapor pressure in winter season |
| W_ cyclone | Cyclone form Yes/ No in Winter Season |
| S_Temp | Temperature of Hot Summer Season in March, April and May month |
| S_Cloud cover | Cloud cover of Hot Summer Season |
| S_Vapor pressure | Vapor pressure of summer season |
| S_ cyclone | Cyclone form Yes/ No in Summer season |
| SW_Temp | Temperature of South west Monsoon in June, July, August and September month |
| SW_Cloud cover | Cloud cover of South west Monsoon |
| SW_Vapor pressure | Vapor pressure of South west Monsoon |
| SW_ cyclone | Cyclone form Yes/ No in Southwest Monsoon |
| NE_Temp | Temperature of North east Monscon in October, November and December month |
| NE_Cloud cover | Cloud cover of North east Monsoon |
| NE_Vapor pressure | Vapor pressure of North east Monsoon |
| NE_ cyclone | Cyclone form Yes/ No in North East Monsoon |

### *A. Random Forest Algorithm*

A random forest is a collection of unpruned decision trees [4,7,8,10]. It combines many tree predictors, where each tree depends on the values of a random vector sampled independently. In order to construct a tree, assume that 'N' is the number of training observations and "S" is the number of attributes in a training set. In order to determine the decision node at a tree, choose N<<S as the number of variables to be selected. Select a bootstrap sample from the N observations in the training set and use the rest of the observations to estimate the error of the tree in the testing phase. Randomly choose m variables as a decision at a certain node in the tree and calculate the best split based on the m variables in the training set. Trees are always grown and never pruned compared to other tree algorithms.

### *B. C4.5 Algorithm*

C4.5 is a decision tree technique which is enhanced by ID3 algorithm. It is one of the most popular algorithm for rule base classification [4,11,13,15]. Here an attributes can be split into two partition based on the selected threshold value, all the value satisfied by the constraint it will be assigned in one child and remaining values can be store in another child respectively. It also handles missing values. Here it can be gather of all nominal tests through entropy gain and the values are sorted based on the values in continuous attribute values which are calculated in one scan. This process is repeated for each continuous attributes when the process is terminated.

Steps of the System:

1. Selecting dataset as an input to the algorithm for processing.
2. Selecting the classifiers.
3. Calculate entropy, information gain, gain ratio of attributes.
4. Processing the given input dataset according to the defined algorithm of C4.5 data mining.
5. According to the defined algorithm of improved C4.5 data mining processing the given input dataset.
6. The data which should be inputted to the tree generation mechanism is given by the C4.5 and improved C4.5 processors. Tree generator generates the tree for C4.5 and improved C4.5 decision tree algorithm

The rule set is formed from the initial state of decision tree. Each path from the initial state, the condition will be evaluate and simplified by the effect of rule and an outcomes will put on the required leaf, the step will continuous when it comes discarding the condition. Let freq ($C_i$, S) stand for the number of samples in S that belong to class $C_i$ (out of k possible classes), and $|S|$ denotes the number of samples in the set S. Then the entropy of the set of equation 1 such as

$$\text{Info}(s)=\sum_{i=1}^{k}((\text{freq}(c_i,s)/|s|).\log_2(\text{freq}(c_i,s)/|s|)) \quad (1)$$

After set T has been partitioned in accordance with n outcomes of one attribute test X by equation 2 and 3,

$$\text{Info}_X(S)=\sum_{i=1}^{n}\frac{|S_j|}{|S|}.\text{Info}\left(S_j\right) \quad (2)$$

$$\text{Gain}(x) = \text{Info}(S) - \text{Info}_x(S) \quad (3)$$

The gain ratio "normalizes" the information gain as following equation 4,

$$\text{GainRatio}(a_i,S)=\frac{\text{InformationGain}(a_i,S)}{\text{Entropy}(a_i,S)} \quad (4)$$

Pre-pruning involves deciding when to stop developing sub-trees during the tree building process.

### *C. C4.5 Algorithm with Bootstrap Aggregation (Bagging) Algorithm*

In third contribution, to integrate C4.5 algorithm combines with Bagging improves generalization error by reducing the variance of the base classifiers [11,12,14]. The performance of bagging depends on the stability of the base classifier. After training the x classifiers, a test instance is assigned to the class that receives the highest number of votes.

Input: D, Set of S training tuples;

Classification learning scheme: C4.5 Algorithm

 

Output: The ensemble- a composite model, M*.
Algorithm: Bagging
Let w be the number of bootstrap samples
**for** i=1 to w **do** // create w models:
Create a bootstrap sample of size S, $D_i$ by sampling D with replacement;
Use $D_i$ and learning scheme to derive a Model, $M_i$;
**endfor**
To use the ensemble to classify a tuple X:
Let each of the w models classify X and return majority vote;

It increases accuracy because the composite model reduces the variance of the individual classifiers.

## III. EXPERIMENT ANALYSIS

In experiment analysis[1], the True Positive rate (TP rate), False Positive rate (FP), Precision, Recall, F-Measure, ROC Area classified based on Class can measured for all algorithm. In bagging with C4.5 by pruned data and its detailed classifying accuracy 99.7% as shown in Table-2 and unpruned of random forest is 97.6% and C4.5 with Bootstrap Aggregation acquired 100% for selective variable as shown in table 2.

Table.2 Detailed Accuracy by Class of Bagging with C4.5

| Algorithm | All Variable (%) | Selected Variables (%) | Sensitivity (%) | Specificity (%) |
|---|---|---|---|---|
| Random Forest (Unpruned) | 97.6 | 98.8 | 93.75 | 98.25 |
| C4.5 (Pruned) | 99.7 | 99.9 | 97.6 | 99.8 |
| C4.5 with Bootstrap Aggregation | 99.8 | 99.96 | 99.7 | 99.9 |

Table 2 represents the sensitivity is the probability that a test will indicate weather climate among those with the weather data.

- **Sensitivity: True positive/(True positive + False Negative) × 100**

Specificity is the fraction of those without disease who will have a negative test result:

- **Specificity: True Negative/(True Negative + False Positive) × 100**

Here Sensitivity and specificity are characteristics of the test. The selected variables alone were used to find the sensitivity and specificity of the data mining algorithms. In C4.5 with Bootstrap Aggregation gives high accurate classification rate than C4.5 pruned and random forest algorithm. When compared with C4.5 pruned and C4.5 with Bootstrap aggregation, it can be slightly vary for observing to determine the size of the tree.

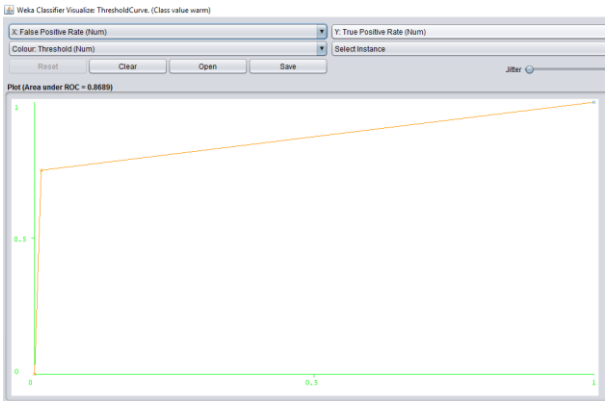


Figure 2. Threshold curve of Roc =0.8689

From the given figure-2 illustrates, the threshold curve has been show the classification accuracy and the value of ROC area appear to be high as 0.8689 , which can be applied in virtual screening and perform cost benefit analysis efficiently.

From the figure-3 From the given figure illustrates, the threshold curve has been show the classification accuracy and the value of ROC area appear to be high as 0.918 , which can be applied in virtual screening and perform cost benefit analysis efficiently.

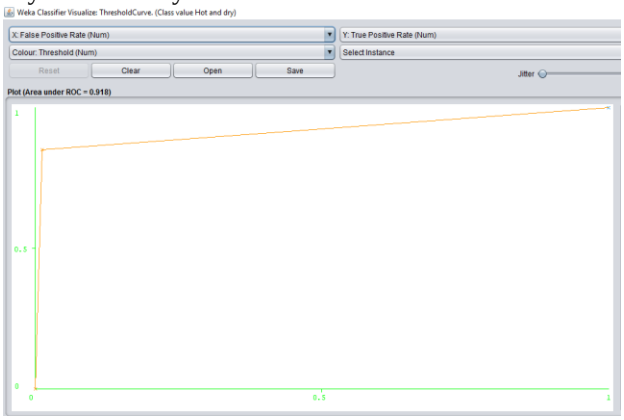


Figure 3. Threshold curve of Roc =0.918

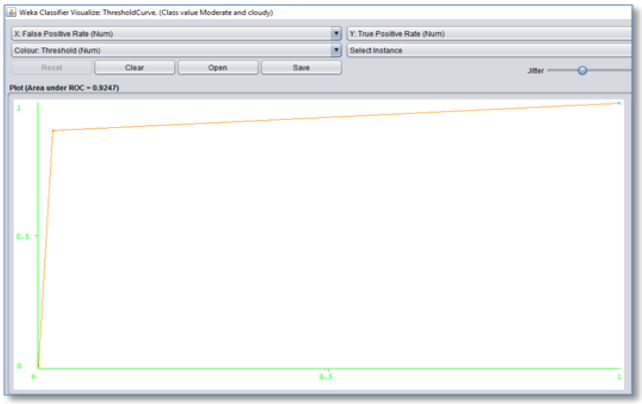


Figure-4 Threshold curve of Roc =0.9247

 

From the figure-4 illustrates, the threshold curve has been show the classification accuracy and the value of ROC area appear to be high as 0.9247hich can be applied in virtual screening and perform cost benefit analysis efficiently.

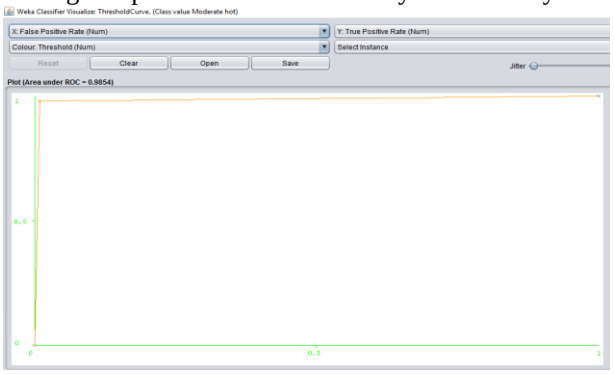

Figure 4. Threshold curve of Roc =0.9854

From the figure-4 illustrates, the threshold curve has been show the classification accuracy and the value of ROC area appear to be high as 0.9854 , which can be applied in virtual screening and perform cost benefit analysis efficiently.

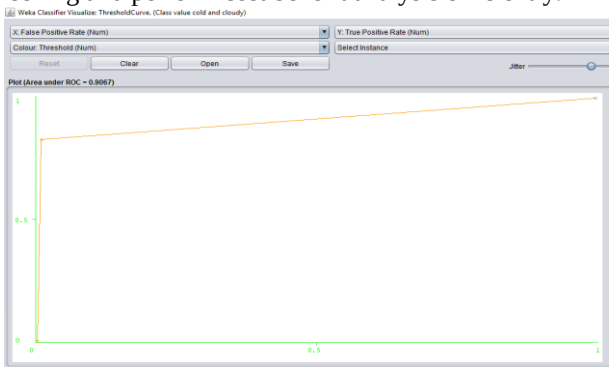

Figure 5 Threshold curve of Roc =0.9247

From the figure 5 illustrates, the threshold curve has been show the classification accuracy and the value of ROC area appear to be high as 0.9067, which can be applied in virtual screening and perform cost benefit analysis efficiently.

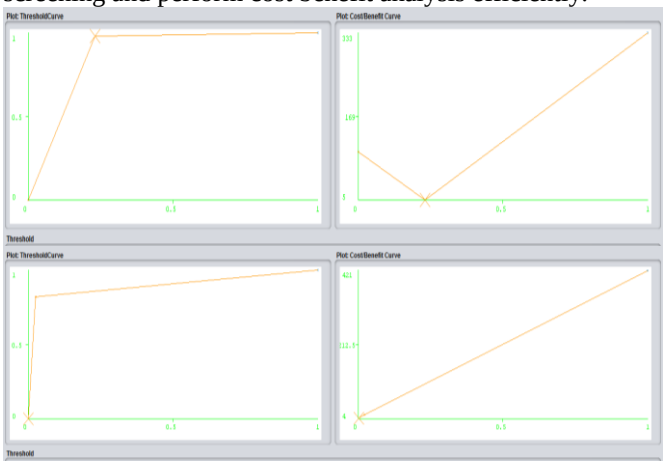

Figure-6 Cost Benefit Analysis

From figure-6 represents, very good Cost Benefit parameters are observed. The Cost is rather low (0), the Gain is rather high (100), in order to retrieve 100 % of "active" ones.

During an implementation, the running time of an executing results taken as 0.12 second in C4.5 Bootstrap aggregation, 0.23 seconds taken by using C4.5 algorithm and 0.3 seconds taken by using random forest algorithm as shown in table-3. By these time efficiency C4.5 with Bootstrap Aggregation takes less time for execution as shown in figure7.

The result suggested that the ensemble method of C4.5 with Bootstrap Aggregation could derive a better classify model in practical as follows,

| | | | |
|---|---|---|---|
| Correctly Classified Instances | 432 | 100 | % |
| Incorrectly Classified Instances | 0 | 0 | % |
| Kappa statistic | 1 | | |

Information Score 435.7428 bits 1.0087 bits/instance

Class complexity | order 0 438.1124 bits 1.0141 bits/instance

Class complexity | scheme 2.3695 bits 0.0055 bits/instance

Complexity improvement 435.7428 bits 1.0087 bits/instance

| | |
|---|---|
| Mean absolute error | 0.0024 |
| Root mean squared error | 0.017 |
| Relative absolute error | 0.888 % |
| Root relative squared error | 4.6602 % |
| Total Number of Instances | 432 |

Table 3. Time efficiency of classification process

| | Random Forest | C4.5 | C4.5 with Bootstrap Algorithm |
|---|---|---|---|
| **Time Taken in Seconds** | 0.3 | 0.23 | 0.12 |

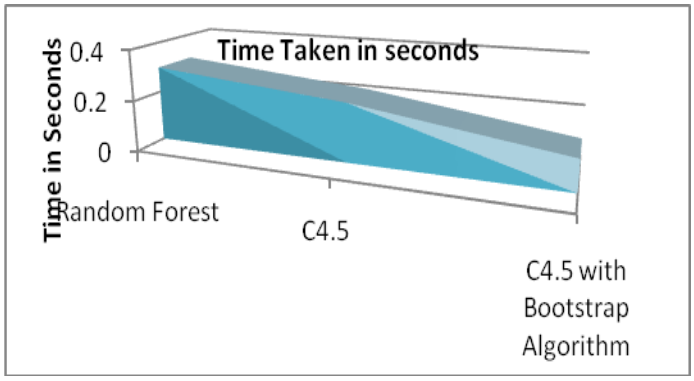


Figure.7 Time taken for Execution

 

## IV. CONCLUSION

From this research, it can be concluded that to evaluate an accurate method of machine learning technique by applying weather data with three contributions. From the research, C4.5 with Bootstrap Aggregation gives high accurate classification rate and taken less time for execution during run time rather than C4.5 pruned and random forest algorithm. Through this research work is contribute to our Government is driving profitability for Regions can be evacuated if hurricanes or floods are expected. Having a stable climate and weather patterns benefits national security. It could also lead to international conflicts as the need for humanitarian aid rises in response to a higher demand on resources. In future, it has been enhanced to analyse big data in various resources.

## REFERENCES


[1] Hall, Mark, et al. "The WEKA data mining software: an update." ACM SIGKDD Explorations Newsletter 11.1 (2009): 10-18.

[2] T.F. Gonzales. "Clustering to minimize the maximum inter cluster distance". Theoretical Computer Science,1985,38(2-3):293-306.

[3] Kannan, M., S. Prabhakaran, and P. Ramachandran. "Rainfall forecasting using data mining technique."(2010)

[4] Arun K Pujari, "Data mining techniques", University Press (India). 2003.

[5] Jiawei Han Micheline Kamber, "Data Mining: Concepts and Techniques", Morgan Kaufmann Publisher an imprint of Elsevier, 2006.

[6] L. Breiman, J. Friedman, R. Olshen and C. Stone. "Classification and Regression Trees", Wadsworth International Group, Belmont, CA, 1984.

[7] Quinlan, J.R.. "C5.0 Online Tutorial", (2003)

[8] Wu, X., Kumar, V., Quinlan, J. R., Ghosh, J., Yang, Q., Motoda, Z., Steinbach,M., Hand, D. J and Steinberg, D (2008). "Top 10 Algorithms in Data Mining", Knowledge and Information Systems, 14 (1): 1-37.

[9] Schapire, R. "The strength of weak learnability", Machine Learning,(1990) 5(2): 197-227.

[10] Breiman, L . "Random Forests". Machine Learning 45 (1): 5–32. (2010)

[11] Freund, Y. Schapire, R. "Experiments with a new boosting algorithm", In Proceedings of the Thirteenth International Conference on Machine Learning, 148-156 Bari, Italy. (1996)

[12] Dietterich, T. G.. "An experimental comparison of three methods for constructing ensembles of decision trees: bagging, boosting and randomization". Machine learning, 40: 139-157. (2000).

[13] Opitz, D and Maclin, R "Popular Ensembl Methods: An Empirical Study", 11: 169-198. (1999)

[14] Quinlan, J. R. "Bagging, Boosting and C4.5", AAAI/IAAI, 1: 725-730. (1996)

[15] M. Mayilvaganan, D. Kalpanadevi, "Comparison of Classification Techniques for predicting the performance of Students Academic Environment" in (2014)


## AURTHOR PROFILE

**Mr. M. Manikandan,** pursuing P.hD in Computer Science in PG and Research department of Computer Science, Marudupandiyasr college of arts and Science, Thanjavur is Under affiliated by Bharathidasan University, Trichy. He complete M.Phil in Computer Ssscience in SASTRA univerisyt at 2010. Also he has 6 years experience in teaching field.

**Second Aurthor:**

**Dr. R. Mala**, pursed M.C.A, M.Phil., Ph.D in Computer Science. She has 13 yerars experience in Teaching field and more than 6 years experience in research field. She has published more the 30 papers in National and International journals. She is working as Assistant Professor and Head incharge of Deparatment of Computer science, Alagappa University college, Paramakudi. She is an efficent and motivated person for the research scholars under her guidance.

 