Random Forest Advice for Water Quality Prediction in the Regions of Kadapa District

S.V.S. Ganga Devi

Abstract— Water is essential to all basic needs of Human being. The quality of water is significant on the earth for everyone. Machine learning methods concentrates much on data rather than methods. Classification technique uses the past history data to predict the class of new sample(s). The present work collects water samples in the regions of Kadapa district, Andhra Pradesh. Those samples are given to the Laboratory to perform an analysis on physico- chemical properties of ground water, whether they are suitable for drinking or not. In this paper, Random Forest approach is considered to predict the water quality in the regions of Study area and classify the regions into 3 classes whether they are Excellent, Good or Poor for drinking purposes.

Key Words— Ground water, Conditional Interference tree, Random Forest model

1. INTRODUCTION

Water is prime natural resource. In this modern environment, fresh water is not available due to increase of population, agricultural and industries.

The quality of water is determined by the physico-chemical properties of water and micro biological characteristics. The physico-chemical analysis of water samples was done my several researchers using by standard methods [7],[8],[9],[11],[13],[17]. Water quality prediction is done using various data mining techniques [15]. In this research [12] data mining techniques are used to study various classifiers and to find out most accurate classifier.

The paper is organized as follows. Section 2 deals with Conditional Inference tree and Random Forest approach for assessing the water quality. Section 3 explains about the results obtained. Section 4 deals with conclusion.

2. MATERIALS AND METHODS RESULTS

Random Forest method considers many classification trees and then collects the predictions from all the trees. Random Forest approach implies variable importance to find the smallest set of predictor variables [6].

In Random Forest approach ensemble of trees is considered rather than a single one [1]. Random forest trees generate the classification tree based on the predictor variables.

First, an analysis on physico chemical properties of ground water quality assessment is done in the laboratory for all the water samples collected from the regions of Kadapa district. Water Quality Index (WQI) values were calculated [4]. After, water quality classification is determined by using WQI values [2]. By using this trained data, Random Forest approach is applied to test the water quality whether they are suitable for drinking or not. R programming code is written for the proposed method and results were displayed.

3. RESULTS AND DISCUSSION

In this research work, Conditional Inference Tree, Random Forest are considered to classify a new tuple. 70% of data is used for training purpose and 30% of data is used as test data for prediction.

3.1 Conditional Inference Tree

Conditional inference trees are similar to the classical decision trees but variables and splits are selected based on significance tests rather than purity/homogeneity measure[16].The significance tests are permutation tests. The output of the model is,

Conditional inference tree with 3 terminal nodes
Response: Class
Inputs: pH, EC, TH, ca, cl, F, TDS
Number of observations: 41

1) TDS <= 1152; criterion = 1, statistic = 21.099
2) EC <= 600; criterion = 0.996, statistic = 15.099
3) * weights = 15
2) EC > 600
4) * weights = 19
1) TDS > 1152
5) * weights = 7

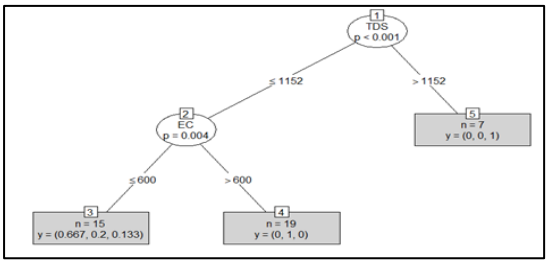

Figure1: Conditional Inference Tree (simple)

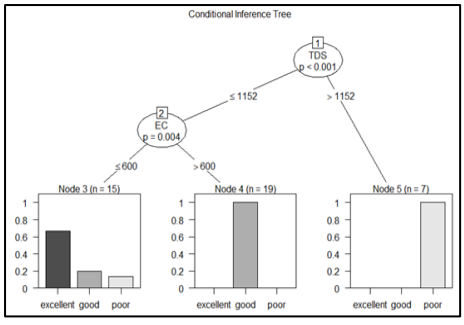

Figure 2: Conditional Inference Tree

Revised Manuscript Received on April 12, 2019.

V.S. Ganga Devi, Madanapalle Institute of Technology & Science, Madanapalle, A.P., India. (E-mail: gangadevisvs@gmail.com)

Conditional inference tree for water quality data consists of 5 nodes as shown in Figure 1 and Figure 2. All the terminal nodes contain the probabilities of each class (Excellent, Good and Poor). For example, terminal node 4 contains training samples 19 and all belong to the class good.

3.2 Random Forest method

In Random forest method the results are aggregated to improve classification rate in Random Forest [3]. The provision of OOB (Out Of Bag) error rates and measures of variable importance are also an important advantage.

Table 1: Parameter and its Gini Vaue

Parameter	Gini value
pH	1.235269
EC	6.500007
TH	1.649205
Ca	1.735817
Cl	3.700561
F	2.384376
TDS	6.942977

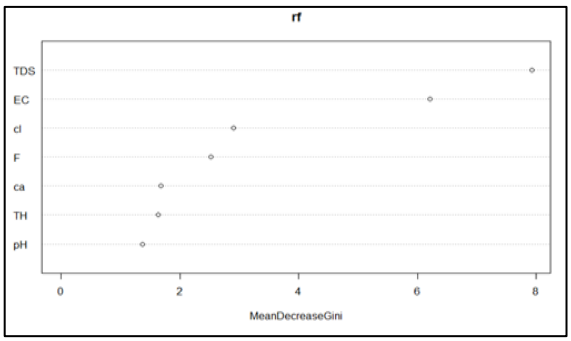

Figure 3: Variable importance

Random forests can provide a measure of variable importance. Node impurity is measured with the Gini coefficient. From Table 1 and Figure 3, it is Known that TDS is the most important variable (which has highest Gini value) and pH is the least importance variable (lowest Gini value). The Confusion matrix (CM) obtained for Random Forest model is displayed in Table 2.

Table 2: CM for Random Forest

Predicted / Actual	Excellent	Good	Poor
Excellent	4	1	0
Good	0	7	0
Poor	0	0	4

Confusion Matrix is a tool in determining the classifier accuracy [5]**.** In Confusion Matrix rows correspond to actual values and columns correspond to predicted ones. An entry in Confusion matrix, CM(i,j) is the number of samples that actually belong to class i, but according to the classifier they are labelled as class j. Among 16 samples of test data, 15 samples are classified correctly and 1 sample is incorrectly classified as shown by Table4, that sample actually belongs to the Excellent category but according to classifier it was labelled as Good. Accuracy is 93.75%.

Random Forest function is used to grow 500 traditional decision trees as shown in Figure 4. The Error is dropping as we keep on adding more and more trees and average them. The mean squared error of the model changes over time as we construct more and more decision trees on random samples of training data. The error rate does not seem to change after a while and there is no sense in generating more than given number of samples. The OOB (Out of Bag) error rate value obtained is 6.25

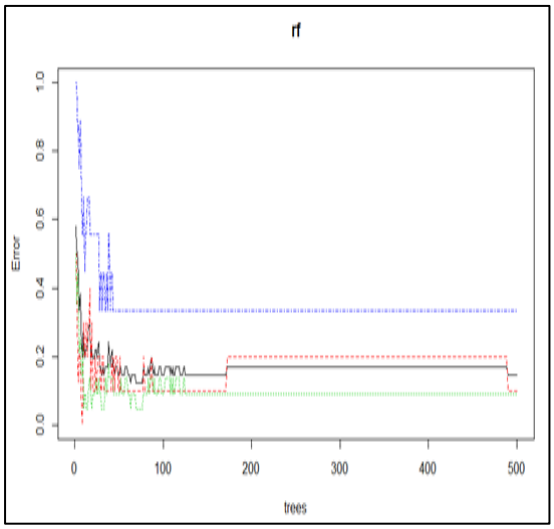

Figure 4: OOB Error rate

4. CONCLUSION

Random Forest approach is one of the best classification approach in Machine Learning. Traditional Decision trees are very large and can be pruned whereas Conditional inference trees are not pruned and accuracy is increased. In Study area, Total Dissolved Solids (TDS) is the most significant variable and pH is the less significant variable according to the variable importance of Random Forest approach. By classifying the given regions into three classes (Excellent, Good, Poor) for drinking purpose, this model has 6.25% as an error rate and accuracy is of 93.75%.

ACKNOWLEDGEMENT

This research work was supported by UGC-SERO for providing the financial assistance to the Minor Research Project [Grant Number: F MRP – 6543/16(SERO/UGC)].

REFERENCES

1. Biau , Gerard (2012), “Analysis of a Random Forests Model” , *The Journal of Machine Learning Research* , No.1 ,pp 1063-1095.
2. Boateng TK, OpokuF, Acquaah So,Akoto O (2016), “Ground water quality assessment using statistical approach and water quality index in Ejsujuaben Municipality,Ghana” ,*Environmental Earth Sciences*, pp 75-489.
3. Breiman, Leo (2001) , “Random Forests” *Machine Learning*, No. 1 , 5–32.

4. Farhad Howladar, Md Abdullah, AI Numanbakth, Mohammad Omar Faruque (2017), "An application of Water quality index (WQI) and multivariate statistics to evaluate the water quality around Maddhapara Granite Mining Industrial area,Dinajpur,Bangladesh" ,*Environmental System Research, Springer Open*.
5. Hamilton, Howard (2012) , "Confusion Matrix", *Knowledge Discovery in Databases*.
6. Hastie, Trevor, Robert Tibshirani and Jerome Friedman (2009) ,*The Elements of Statistical Learning: Data Mining, Inference, and Prediction* ,NewYork: Springer.
7. J. Lu, T. Huang(2009) , "Data Mining on Forecast Raw Water Quality from Online Monitoring Station Based on Decision making Tree", *Fifth International Joint Conference on INC, IMS and IDC.*
8. Jorge Camejo, Osvaldo Pacheco, Miguel Guevara (2013), "Classifier for Drinking water Quality in real time", *Foundation for Science and Technology, IEEE*.
9. M.J. Diamantopoulou, V.Z. Antonopoulos and D.M. Papamichail (2005) , "The use of a Neural Network technique for the prediction of water quality parameters of Axios River in Northern Greece" , *EuropeanWater,*11/12, pp 55-62 .