Enhanced Gradient Boosting Regression Tree For Crop Yield Prediction

K.Shyamala, I.Rajeshwari

Abstract— Agriculture, the main occupation and backbone of our country, is one of the most important fields in the emerging real world and is in poor condition due to the lack of proper guidance to the farmers. This work presents an approach which uses modified gradient boosting regression technique to predict the yield of the crops to be cultivated based on the weather condition and the season. This is done by applying different statistical techniques in computing the minimum weight of the leaf, minimum samples for split and least squares error. The dataset has been collected from the publicly available Indian Government Records. From the dataset, two datasets of size 2000 and 4000 are formed. The original and modified algorithm were compared based on the metrics accuracy on training set, accuracy on test set, mean accuracy and standard deviation, MAE (Mean Absolute Error), MSE(Mean Squared Error) and R squared score by applying them on two datasets. The modified algorithm shows a better result.

Index Terms— Agriculture, Data Mining, Decision tree, Gradient Boosting, MAE, MSE, Yield prediction.

1. Introduction

Agriculture is the backbone of Indian Economy. More than half of population is dependent on agriculture. The agriculture output is very low nowadays. But the demand for food is increasing. Hence there is need for initiative steps to face the demand. The farmers, agricultural scientists and government are finding out ways to put extra effort and techniques for more production. Crop yield prediction is done by the farmers traditionally out of their own experience on their field and crop. Data mining is a process of identifying previously unknown inferences from the huge volume of available data. More researches are being carried out by using data mining techniques in agriculture. This paper focuses on yield prediction which is a major problem, our farmers are facing now. This work is carried out on the dataset consisting of weather and season data along with the crop and the corresponding yield in ton. The data pertaining to 20 years, month wise and season wise are used. The data have been collected through internet from the publicly available government data site[1][2]. The main aim of the paper is to assist farmers in crop selection for cultivation based on the predicted yield, the current season and climate. From the collected dataset, two datasets of size 2000 and 4000 are formed. The entire dataset tuples are split into two sets., viz., a training set and test set. The training sets are randomly sampled from the dataset. The remaining tuples form the test set and are independent of the training tuples, meaning that they will not be used to build the regressor [3]. The gradient boosting regression trees are constructed using the collected datasets. Then the modified gradient boosting regression algorithm (MGBR) is applied on the same datasets. Their performance are analysed based on certain metrics like Accuracy score, Mean Absolute Error, Mean Squared Error, R2 score and execution time.

- *Dr. K. Shyamala, Associate Professor of Computer Science, Dr. Ambedkar Govt. Arts College, Chennai, Tamilnadu, India. shyamalakannan2000@gmail.com*
- *Mrs. I. Rajeshwari, Research Scholar, Department of Computer Science, Dr. Ambedkar Govt. Arts College, Chennai, Tamilnadu, India. rajeshwari_i@yahoo.com*

The organization of this paper is as follows. In Section 2, the literature review conducted is described and the study to be done is introduced. Section 3 describes the dataset collection, its preprocessing activities and the regression tree algorithms used. Section 4 discusses the results of the algorithms for predicting the crop yield. Finally, section 5 includes conclusion and future work.

2 Literature Review

Branko et al. [4] had used M5P model tree for yield prediction. They had taken maize, soybean, sugar beet and other field cultures. The attributes considered were maximal, minimal and average monthly air temperature, precipitation level, etc. About 41 attributes were initially considered and after applying feature subset selection methods, the dimensionality was reduced and the performance was improved. Ramesh et al. [5], in their paper, had used the dataset taken for the years 1965 to 2009 in East Godavari district of Andhra Pradesh in India. The attributes considered were Year, Rainfall, Area of Sowing and Production. MLR technique and K-Means algorithm were applied on the dataset for yield prediction. Based on accuracy, MLR technique outperformed the K-Means algorithm. Zingade et al. [6], had proposed a system which takes the location of the user as an input, from which, the soil nutrients such as Nitrogen, Phosphorous, Potassium, forecasted weather are obtained. The Multiple Linear Regression algorithm is applied on the resultant dataset to propose the best feasible crops according to given environmental conditions. Considering the past year production, the expected yield for the crop proposed is given. Sellam et al. [7], had used, Linear Regression Analysis to establish a relationship among a set of variables Annual Rainfall, Area under Cultivation and Food Price Index and their effects on yield of rice crop. Data pertaining to 10 years had been taken. Poongodi et al. [8], had used rainfall, soil data and climate dataset to predict the crop production. Optimal features are selected using genetic algorithm. Then improved C4.5 in the hidden layer with ANFIS classifier is used for classifying the data based on region wise. C4.5 classifier generates the rules to predict the crop yield. Based on accuracy, the performance of this model is better when compared with the existing classifier. Noronha et al. [9], had analysed the various data mining techniques that are in use for the crop yield prediction. It includes K-Means, K-Nearest neighbor (KNN), Support Vector Machine (SVM), Multiple Linear Regression (MLR) and Biclustering techniques. They

have concluded that Biclustering techniques are rarely used in crop yield prediction and have a major scope to analyze when compared to other clustering techniques. Shilpa et al. [10], had implemented different clustering methods for the districts having similar kind of productivity factors for crops and compared their performance; and implemented regression analysis for forecasting the yield of the major crops for different districts. The dataset consists of 13 attributes like area, production, temperature, rainfall, pH, Soil minerals (Nitrogen, phosphorus, potassium) etc. They had concluded that the effectiveness of clustering algorithm is data dependent for regression analysis, if the data collected is significant then the results will fit the model, otherwise it can lead to some imprecise results. Devika et. al. [11], had used the attributes year, crop, area and production (in tons) in the dataset and applied Linear regression algorithm on it. The model was validated based on the metrics Multiple R square, Adjusted R squared and F-Statistic values. The authors have suggested linear regression for Ecuadorian conditions and concluded that using the system the yield of sugarcane, cotton, and turmeric are predicted in the highest level. Veenadhari et. al. [12], had used ID3 algorithm for predicting the soybean yield in Bhopal district. They had used climatic data pertaining to the period 1984-2003. The attributes taken into consideration are average rainfall, average evaporation, average maximum temperature, average maximum relative humidity and average soybean yield. On this dataset, ID3 algorithm was applied. They have concluded that the analysis of decision tree indicated that the Relative humidity has the highest influence on soybean crop productivity followed by temperature and rainfall. From the review, it is found that different datasets with different attributes are being used for yield prediction. Various data mining algorithms are used and most of the authors have analysed them based on few metrics. In this study, gradient boosting regression algorithm and the MGBR algorithm are applied on the two different size datasets and analysed on more metrics like accuracy on training set, accuracy on test set, mean accuracy and standard deviation, MAE (Mean Absolute Error), MSE(Mean Squared Error) and R squared score.

3 Research Methodology

3.1 Data collection, cleaning and coding:

From the publicly available Indian Government Records, the dataset has been collected. The attributes included are year, month, average temperature, minimum temperature, maximum temperature, diumal temperature, cloudcover, potential, evapotranspiration, reference crop evaporation, precipitation, vapour pressure, wet day frequency, rainfall, season, crop and yield. The collected data was preprocessed by removing unwanted data, noisy data and blank data. Extreme values in each attributes are treated as noisy data. The attribute year and month are removed. The data was initially coded in EXCEL and finally converted to comma delimited .csv. The data were taken as two datasets with sizes 2000 and 4000.

- The minimum weighted fraction of a leaf node is the minimum weighted fraction of the sum total of weights (of all the input samples) required to be at a leaf node. In the gradient boosting algorithm, when this minimum weighted fraction of the leaf node is not zero and the sample weight is provided the minimum weight of the leaf is calculated as the product of minimum weighted fraction of the leaf node and sum of all sample weights.

In the MGBR algorithm, the minimum weight of the leaf (MWL) is calculated as the product of minimum weighted fraction of the leaf node (MWFL) and n percentile of all sample weights (SW). The n is equal to alpha *100.

3.2 Decision tree regressor:

As the class is a continuous one, regression is used [13]. Decision tree regression observes features of an object and trains a model in the structure of a tree to predict data in the future to produce meaningful continuous output [14]. Continuous output means that the output/result is not discrete, i.e., it is not represented just by a discrete, known set of numbers or values.

3.3 Gradient Boosting:

In applied machine learning, one of the most powerful booming techniques is Gradient boosting. Boosting is based on the ensemble principle. It is a sequential technique that combines a set of weak learners and delivers improved prediction accuracy. At any time instant, say, I, the outcomes of the model are weighed based on the outcomes of previous instant I-1. A lower weight is given to the correctly predicted outcome and a higher weight for the misclassified one [15].

3.4 Modified Gradient Boosting Regression (MGBR):

In the MGBR, the following statistical methods are applied over traditional gradient boosting algorithm. In the decision tree construction, the minimum number of samples required to split an internal node is taken as the maximum of minimum samples split and 2 times the minimum number of samples required to be at a leaf node. A split point at any depth will only be considered if it leaves at least minimum number of samples required to be at a leaf node as training samples in each of the left and right branches. In the MGBR algorithm, the minimum number of samples required to split an internal node (MSS) is taken as the maximum of minimum samples split and 4 times the minimum number of samples required to be at a leaf node(MSL).

$$\boldsymbol{MSS = Max(MSS, 4 * MSL)}$$

$$\boldsymbol{MWL = MWFL * percentile(SW, n)}$$

When calculating the least square loss function in the Gradient Boosting Regression, if the sample weight is not null, the error is calculated as

$$\boldsymbol{LSE = \frac{1}{Sum(SW)} * \sum SW * (y - predictedy)^2}$$

In the MGBR algorithm, it is calculated as

$$\boldsymbol{LSE = \frac{1}{Mean(SW)} * Mean(SW * (y - predictedy)^2)}$$

Algorithm : Modified Gradient Boosting Regression (MGBR)
Input : Weather Dataset

Output : Decision Tree Regressor

```
Begin
    if Minimum samples split is in integer then
        If x < 2 then
            Return an error that the Minimum samples split should be greater than 1
        End if
    Else // Minimum samples split is in float
        If Minimum samples split is not in the range (0.0, 1.0) then
            Return an error that the Minimum samples split should be a value between 0.0 and 1.0
        End if
        Minimum samples split = Maximum (2, integer(ceiling(Minimum samples spilt * no. of samples)))
    End if
    Minimum samples split = Maximum ( Minimum samples split , 4 * Minimum samples leaf)
    if Sample Weight = None OR Minimum weight fraction leaf = 0 then
        Minimum weight leaf = 0
    Else
        Minimum weight leaf = Minimum weight fraction leaf * Percentile (Sample weight, alpha * 100)
    End if
    if the sample weight is null then
        Return the mean value of (y – predictedy)^2
    Else
        Return the result of 1/mean value of sample weight * mean value of (sample weight * (y – predictedy)^2)
    End if
    End
```

4 Result And Discussions

The algorithms were executed on the processed datasets in .CSV (Comma Separated format) file using PYTHON and the various metrics were analysed. Table 4.1. shows accuracy on training set, accuracy on test set, mean accuracy and standard deviation resulted by the algorithms applied on the processed datasets with size 2000 and 4000. Figure 4.1. shows the performance of the algorithms on the datasets based on the mean accuracy. As can be seen from the table 4.1, the MGBR algorithm has more accuracies than the original gradient boosting regression algorithm. Standard deviation accuracy should be a minimum one. The table 4.1. shows that the MGBR algorithm results in the minimum standard deviation accuracy, i.e., it results in more accurate prediction.

Table 4.1. Performance of the algorithms based on various accuracies

ALGORITHM	ACCURACY ON TRAINING SET (ATS)		ACCURACY ON TEST SET (AT)		ACCURACY (MEAN) (AM)		ACCURACY (STD DEV.) (AS)	
Dataset size	2000	4000	2000	4000	2000	4000	2000	4000
GBR	0.998	0.985	0.759	0.688	0.77688	0.53633	0.05604	0.18589
MGBR	0.998	0.986	0.784	0.739	0.77985	0.55566	0.04996	0.14762

Figure 4.1. Performance of the algorithms based on mean accuracy

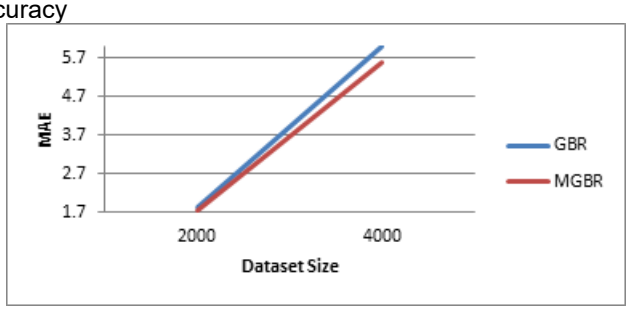
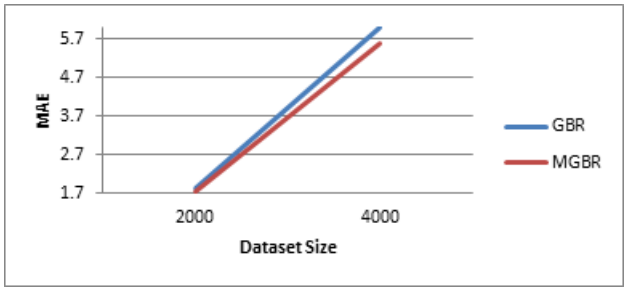


Table 4.2. Performance of the algorithms based on MAE, MSE, R^2 score and execution time

ALGORITHM	MEAN ABSOLUTE ERROR (MAE)		MEAN SQUARED ERROR (MSE)		R^2 SCORE		EXECUTION TIME	
Dataset size	2000	4000	2000	4000	2000	4000	2000	4000
GBR	1.81831	5.99747	29.7434	548.911	0.75923	0.6885	1m 26.3s	3m 17s
MGBR	1.74857	5.57878	26.6529	460.538	0.78424	0.7386	1m 9.2s	2m 16s

Figure 4.2. Performance of the algorithms based on mean absolute error



The table 4.2 shows Mean Absolute Error, Mean Squared Error, R^2 score and the execution time of the algorithms applied on the datasets. Figure 4.2. shows the performance of the algorithms on the datasets based on the mean absolute error. It is clear that MGBR algorithm has the minimum error and execution time than the original algorithm. The following are the outcomes of the study. The MGBR algorithm has an improved accuracy, reduced error and less execution time than gradient boosting regression algorithm.

5 Conclusion And Future Work

The MGBR algorithm has been designed and its performance has been compared with that of original gradient boosting regression algorithm on the weather and season dataset for yield prediction based on various metrics. The MGBR algorithm has a better performance than the original gradient boosting regression algorithm for yield prediction. The work can be extended further by comparing the performance of these algorithms on different datasets.

References

[1] https://www.indiawaterportal.org/met_data/
[2] https://data.gov.in/

[3] "A Comparative Study of Selected Classification Algorithms of Data Mining", Ashish Kumar Dogra, Tanuj Wala, IJCSMC, Vol. 4, Issue. 6, June 2015, pg.220 – 229, ISSN 2320–088X
[4] "Data Mining Approach for Predictive Modeling of Agricultural Yield Data", Branko Marinković, Jovan Crnobarac, Sanja Brdar, Borislav Antić, Goran Jaćimović, Vladimir Crnojević
[5] "Data Mining Techniques and Applications to Agricultural Yield Data", D.Ramesh , B.Vishnu Vardhan, International Journal of Advanced Research in Computer and Communication Engineering, Vol. 2, Issue 9, September 2013
[6] "Machine Learning based Crop Prediction System Using Multi-Linear Regression", D.S. Zingade, Omkar Buchade, Nilesh Mehta, Shubham Ghodekar, Chandan Mehta, International journal of Emerging Technology and Computer Science ISSN: 2455 9954
[7] "Prediction of Crop Yield using Regression Analysis", V. Sellam and E. Poovammal, Indian Journal of Science and Technology, Vol 9(38), 10.17485/ijst/2016/v9i38/91714, October 2016, ISSN (Print) : 0974-6846, ISSN (Online) : 0974-5645
[8] "Prediction of Crop Production using Improved C4.5 with ANFIS Classifier", S. Poongodi and M. Rajesh Babu, International Journal of Control Theory and Applications, ISSN : 0974-5572, International Science Press, Volume 10 • Number 21 • 2017
[9] "Comparative Study of Data Mining Techniques in Crop Yield Prediction", Perpetua Noronha, Divya .J, Shruthi .B.S., IJARCCE International Journal of Advanced Research in Computer and Communication Engineering, ICRITCSA M S Ramaiah Institute of Technology, Bangalore Vol. 5, Special Issue 2, October 2016 Copyright to IJARCCE, DOI 10.17148/IJARCCE 132
[10] "Applying Data Mining Approach and Regression Model to Forecast Annual Yield of Major Crops in Different District of Karnataka", Shilpa Ankalaki, Neeti Chandra, Jharna Majumdar, International Journal of Advanced Research in Computer and Communication Engineering, ICRITCSA M S Ramaiah Institute of Technology, Bangalore Vol. 5, Special Issue 2, October 2016 Copyright to IJARCCE DOI 10.17148/IJARCCE 25
[11] "Analysis of crop yield prediction using data mining technique to predict annual yield of major crops", B. Devika and B. Ananthi, International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056, Volume: 05 Issue: 12 | Dec 2018 www.irjet.net p-ISSN: 2395-0072 ,© 2018, IRJET | Impact Factor value: 7.211 | ISO 9001:2008 Certified Journal | Page 1460
[12] "Soybean Productivity Modelling using Decision Tree Algorithms", S. Veenadhari, Dr. Bharat Mishra and Dr.CD Singh, International Journal of Computer Applications (0975 – 8887), Volume 27– No.7, August 2011
[13] https://math.stackexchange.com/
[14] https://www.geeksforgeeks.org/python-decision-tree-regression-using-sklearn/
[15] https://www.analyticsvidhya.com/blog/2016/02/complete-guide-parameter-tuning-gradient-boosting-gbm-python/