Yield Forecasting for Indian Crops with Ensemble Model

Lokesh C K, Senthil S

Abstract: — In fast developing economies like India, there is fast depletion of agricultural lands and many people migrate to cities leaving agriculture. In these conditions effective yield prediction of crops is a must for planning food policies and securing food for the people. Import only at time of huge price rise is not a solution and predictive import of crops based on yield predicted is a way to keep food inflation under control. Towards this end, an Ensemble model for prediction of yield of India Rabbi Crops is proposed in this work. An Ensemble machine learning model is built based on past histories and macro climatic and monsoon conditions to forecast the yield for crops in work.

Index Terms: Ensemble Model, Machine learning, ARIMA-LR, Bayes net. Kmeans-Neural Network

I. INTRODUCTION

India has the second highest population in the world. With rapid industrialization, agriculture is facing a threat with dwindling agricultural lands and lack of human resources for agricultural works. The food production is rapidly declining and this decrease is further increased with global climatic crisis. In these conditions, proper food policies have to be framed year to year to ensure food security. Crop yield forecasting is an important activity in this process of devising national food security policies. Use of machine learning techniques for predicting crop yield is the scalable and most cost effective way for a bigger country like India where the crop production is dependent on multiple factors. The crop yield in India is mainly dependent on climate, rain conditions. Factors like availability of fertilizers, pesticides, human labor etc do influence the production but these parameters can be controlled. Information about land cover for crop is also a important influencing factor but it varies and cannot be collected every year. In such cases other parameters like past production data must be used to model these parameters. The problem with adoption of machine learning models for predicting crop yield is its reliability in terms of accuracy. Many works has been done for predicting crop yield using machine learning models in the literature. Most of works are based on time series regression or using supervised machine learning algorithm to learn the relation between yield and dependent parameters. The accuracy is relatively low in these approaches. In this work we propose an ensemble model to forecast crop yield. Ensemble models have been used previously in many classification methods to improve the accuracy of classification. By combining the results of multiple classification models, the accuracy gaps with individual classifiers are compensated. In this works, ensemble model is used for forecasting crop yield. Individual classifiers are trained using different parameters to predict crop yield and the results of these classifiers is combined using weighted average ensemble to forecast the crop yield. The weights for ensemble are learnt continuously using the mean square error of individual classifiers.

Revised Manuscript Received on 30 July 2019.
* Correspondence Author

Lokesh C K, *School of CSA*, REVA University, Bengaluru, India.
Dr. S. Senthil , *School of CSA*, REVA University, Bengaluru, India

II. RELATED WORK

An [1] author applied neural network to predict the crop yield. By using PH of the soil, nitrogen level, temperature and rain fall the crop yield is predicted. The method is only for micro level prediction in a farm land and not applicable for wider area forecasting. In [2] authors applied Multiple Linear Regression (MLR) and Clustering with Density based methods for predicting the crop yield. Nitrogen, The result of MLR model is cross verified using Density based clustering. In [3] authors explored the use of Bayesian Networks for crop yield prediction. By using parameters like precipitation, temperature range, evaporation, transpiration, area, production and yield for the Kharif season a Bayesian model is built to predict crop yield. Based on these parameters BayesNet and Naïve Bayes classifier were trained to predict crop yield. The accuracy of BayesNet was found to be higher than Naïve Bayes classifier. The same author applied neural network for rice yield prediction in [4]. A Multilayer Perceptron Neural Network was used to forecast and it achieved an accuracy of 97.5%. Crop yield prediction based on soil quality level using machine learning is done in [5]. KNN and Naïve Baiyes classifier are used for prediction in this work. The work is for selection of crops for soil type to maximize the production. In [6] ANN (Artificial neural network) and MLR (multiple regression models) are employed to predict the yield. The authors applied the model for sesame crop. The plant characters used for prediction were flowering time, the plant height, the capsule number per plant, seed number per capsule. ANN model performed better than MLR model. A micro level crop prediction based on temperature, rain fall, PH of soil, Nutrients in the soil was proposed in [7]. The approach used ANN for prediction of yield. In [8] authors proposed three model for yield prediction. Fuzzy logic (FL), Adaptive Neuro Fuzzy Inference System (ANFIS) and Multiple Linear Regression (MLR) are used.

The parameters used for training the predictor were based on the biomass and souil moisture. Out of all three models, ANFIS model was able to forecast the yield accurately than Fuzzy and MLR models. In [9] authors applied random forest algorithm for crop yield prediction.

Using the soil nutrients contents and climate conditions, the yield is predicted using ID3 decision tree.It is micro level prediction model.

Author in [10] applied four different models of SVM, AdaBoost, MLR and MNR (Modified Non Linear Regression) for predicting rice yield. Their experiment demonstrated that MNR is able to predict yield more accurately than others. Authors in [11] proposed the soil contents and overall climatic conditions to make decision of yield. Clustering of past history of crop production against these parameters are collected and clustering model is used to predict the crop yield. Most solutions in the survey were based on application of individual models and there is no way to improve the accuracy when considering multiple parameters which can influence the crop yield.

III. ENSEMBLE FORECASTING OF CROP YIELD

The proposed solution is based on ensemble of multiple classifiers trained with different parameters influencing crop yield.

A. *ARIMA-LR*

In India most of the agriculture operations is dependent on monsoons. Due to variation is monsoon behavior the yield of the crop also fluctuates. In seasons of high monsoons, the yield too is higher. ARIMA (Auto regressive integrated moving average) model is one of the best to capture the behavior of the seasonality in monsoon. ARIMA model is built based on the past history of monsoon. A Linear regression is built between the monsoon rain level(cm) and yield of the crop. Once the model is built, the rain level is forecasted using ARIMA and the forecast level is feed to Linear Regression model to predict the crop yield. The rain fall data over period of several years are collected and ARIMA(0,1,1)x(0,1,1) model is constructed on it. The forecasting equation in model is given as

$$\check{Y}_t = Y_{t-12} + Y_{t-1} - Y_{t-13} - \theta_1 e_{t-1} - \emptyset_1 e_{t-12} + \theta_1 \emptyset_1 e_{t-13}$$

Where is the MA(1) coefficient and is the SMA(1) coefficient. This model is abstractly similar to the Winters model insofar as it effectively applies exponential flattening to level, trend, and seasonality all at once, although it rests on more solid theoretical fundamentals, particularly with regard to calculating confidence intervals for long-term forecasts. A generalized linear model (GLM) is constructed between the rain fall level in that year and the crop yield for that year. GLM is more suitable for both continuous and categorical predictors.Since yield is continuous, GLM is used in this work. Models like MLR,ANOVA ANVOCA with fixed effects are included in the GLM model. GLM is given as

$$y_i = N(x_i^T \beta, \sigma$$

where known covariates are contained in and coefficients of model are in . Least squares and weighted least squares are used to fit the model.

B. *Baiyes Net*

Baiyes classifier is built with crop yield as the output and Precipitation, average temperature, average humidity, area of cultivation, production cost ratio. Since the parameters have higher dependences, Baiyes model is used rather than Naïve Baiyes. The best Bayesian network for the input dataset is learnt using the MDL (minimal description length) scoring function. The MDL scoring function for the dataset D is given as

$$MDL(B|D) = \frac{\log N}{2}|B| - LL(B|D)$$

Where |B| is the number of parameters in the network and LL is given as

$$LL(B|D) = \sum_{i=1}^{N} \log\left(P_B\left(u_i\right)\right)$$

The function can be rewritten as

$$LL(B|D) = \sum_{i=1}^{N} \log\left(P_B\left(c^i | a_1^i, \dots . a_n^i\right)\right) + \sum_{i=1}^{N} \log\left(P_B\left(a_1^i, \dots . a_n^i\right)\right)$$

Where |B| is the number of parameters in the network and LL is given as The function can be rewritten as The probability of how well the Bayesian estimates the class is given by the first term and the joint distribution is given by the second term.

C. *Kmeans-ANN*

The past history of yield against production cost ratio, area of cultivation, water level ratio in dams and tanks near major production areas are collected and clustered using K-Means. The outliers data who are present in same clusters but very large difference is yields are removed and the processed cluster is used to train a Multi layer Perceptron neural network.

D. *Ensemble*

The entire data set is split to 80:20 ratio and training is done with 80% data set. The forecasting is done using all the three models ARIMA- LR, Baiyes Net, Kmeans – ANN and the Mean Square Error is measured against each classifier. The weight of the each classifier is calculated as

$$w(C_i) = \log \frac{1 - error\,(C_i)}{error\,(C_i)}, 1 \le C_i \le 3$$

The average ensemble of the result is done as

$$R = \frac{w(C_1) * R(C_1) + w(C_2) * R(C_2) + w(C_3) * R(C_3)}{3}$$

IV. EXPERIMENTAL RESULTS AND ANALYSIS

The crop yield for last 20 years is taken from Indian Government crop statistics website [12]. The model is built for different crops like Rice, Wheat, and Sugarcane etc. The solution was able to achieve maximum accuracy of 99% with ensemble classifier.

Retrieval Number: B2220078219/19©BEIESP
DOI: 10.35940/ijrte.B2220.078219
Journal Website: www.ijrte.org

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication

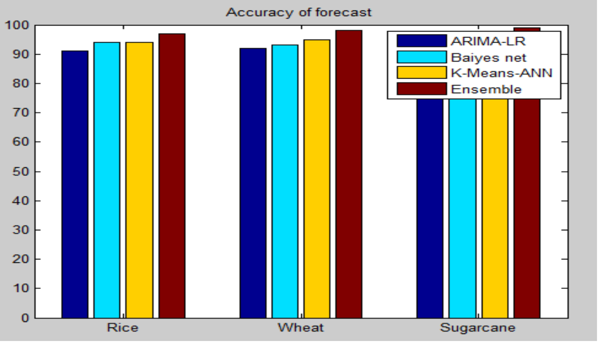

Fig 1: The accuracy values for each crop presented in the above figure.

Table 1: : The accuracy values for each crop is given below table

Crop	ARIMA-LR	Bayes-Net	Kmeans-ANN	Ensemble
Rice	91	94	94	97
Wheat	92	93	95	98
Sugarcane	93	95	97	99

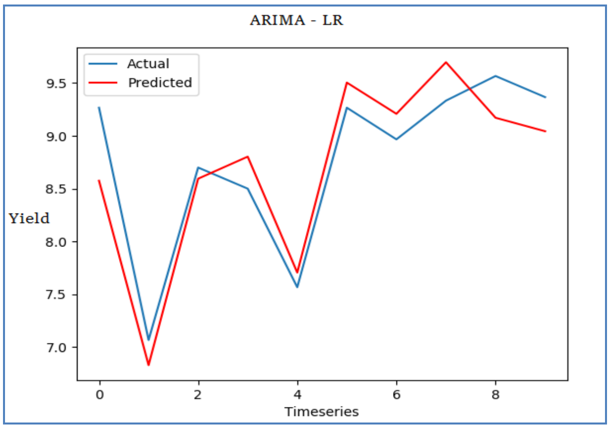

Fig 2: The values of the actual and predicted yield in case of ARIMA-LR are represented using above figure.

Table 2: The values of the actual and predicted yield in case of ARIMA-LR are given below.

Actual	Predicted
9.3	8.6
7.2	6.9
8.7	8.6
8.5	8.8
7.6	7.7
9.3	9.5
9.0	9.2
9.3	9.6
9.5	9.2
9.4	9.0

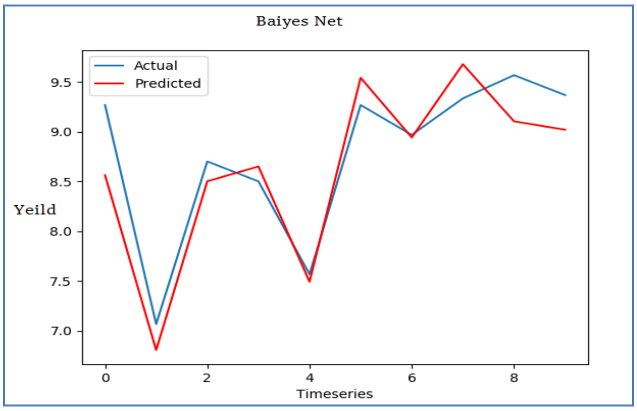

Fig 3: The values of the actual and predicted yield in case of Baiyes net is represented in the above figure.

Table 3: The values of the actual and predicted yield in case of Baiyes net is given below table.

Actual	Predicted
9.3	8.5
7.2	6.6
8.7	8.6
8.5	8.8
7.5	7.7
9.2	9.5
9.0	9.2
9.2	9.6
9.0	9.2
9.4	9.0

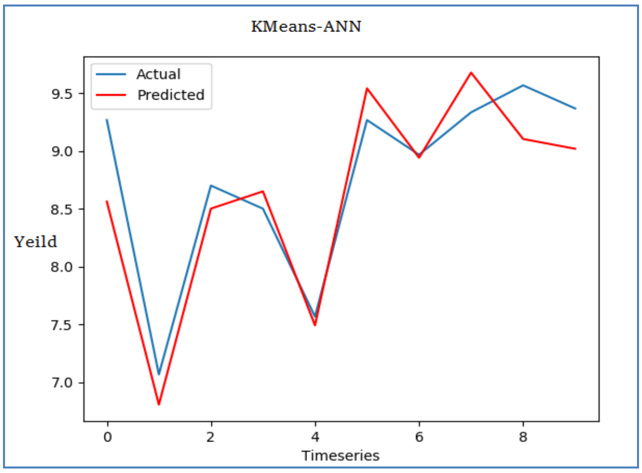

Fig 4: The values of the actual and predicted yield in case of Kmeans-ANN are represented in the above figure.

Table 4: The values of the actual and predicted yield in case of Kmeans-ANN is given the below table.

Actual	Predicted
9.3	8.5
7.2	6.6

8.7	8.5
8.5	8.8
7.5	7.5
9.2	9.5
9.0	9.0
9.2	9.6
9.0	9.2
9.4	9.0

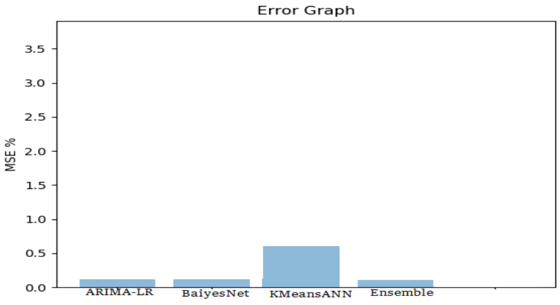

Fig 5: The MSE in each of the classifier is represented in the above figure

Table 5: The values of the actual and predicted yield in case of Ensemble is given below

Method	MSE value
ARIMA-LR	0.2
Baiyes-Net	0.2
Kmeans-ANN	0.5
Ensemble	0.16

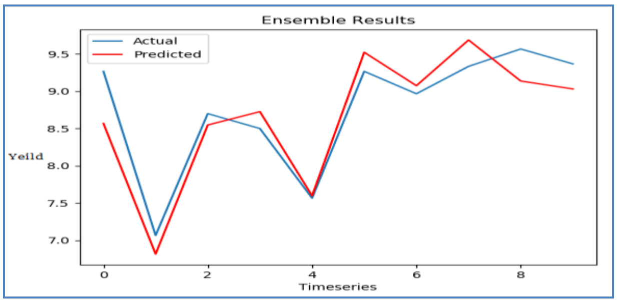

Fig 6: The values of the actual and predicted yield in case of Ensemble are represented in the above figure.

V. CONCLUSION

Ensemble based crop yield forecasting is proposed in this work. Each of machine learning models of ARIMA-LR, Kmeans-Neural Network and Baiyes Net are trained using different parameters and their results are ensemble using weighted average ensembling. The weights for ensemble were adjusted based on the MSE of the individual classifier. Experimental results demonstrate that the accuracy of ensemble classifiers is higher than other individual classifiers.

REFERENCES

1. S S. Dahikar and S. Rode, "Agricultural crop yield prediction using artificial neural network approach", International Journal of Innovative Research in Electrical, Electronics, Instrumentation and Control Engineering, vol. 2, no. 1, pp. 683-686, 2014.
2. D. Ramesh and B. Vardhan, "Analysis of crop yield prediction using data mining techniques", International Journal of Research in Engineering and Technology, vol. 4, no. 1, pp. 47-473, 2015.
3. N.Gandhi, L.J. Armstrong and O. Petkar, "Predicting Rice Crop Yield using Bayesian Networks", communicated, 2016.
4. N.Gandhi, L.J. Armstrong and O. Petkar, "Rice Crop Yield Prediction in India using Artificial Neural Network", International Conference on 2016 IEEE Technological Innovations in ICT for Agriculture and Rural Development (TIAR), Chennai, India scheduled on 15th and 16th July 2016.
5. Monali Paul, Santosh K. Vishwakarma, Ashok Verma. Analysis of soil behaviour and prediction of crop yield using data mining approach. Computational Intelligence and Communication Networks (CICN). 2015; 766-771.
6. Samad Emamgholizadeh, M. Parsaeian, Mehdi Baradaran. Seed yield prediction of sesame using artificial neural network. European Journal of Agronomy. 2015;68, 89-96.
7. Snehal S. Dahikar, Sandeep V. Rode, Pramod Deshmukh, "An Artificial Neural Network Approach for Agricultural Crop Yield Prediction Based on Various Parameters", published in International Journal of Advanced Research in Electronics and Communication Engineering (IJARECE), vol. 4, no. 1, January 2015.
8. Aditya Shastry, H.A. Sanjay, Madhura Hegde. A parameter based ANFIS model for crop yield prediction. Advance Computing Conference (IACC). 2015; 253-257.
9. ShriyaSahu,MeenuChawla,NilayKhare "An efficient analysis of crop yield prediction using Hadoop framework based on random forest approach", 2017 International Conference on Computing, Communication and Automation (ICCCA)
10. Umid Kumar Dey, Abdullah Hasan Masud, Mohammed Nazim Uddin, "Rice yield prediction model using data mining", International Conference on Electrical Computer and Communication Engineering (ECCE), February 16-18, 2017.
11. A. Ahamed, N. Mahmood, N. Hossain, M. Kabir, K. Das, F. Rahman, R. Rahman, "Applying data mining techniques to predict annual yield of major crops and recommend planting different crops in different districts in Bangladesh", 16th IEEE/ACIS International Conference on Software Engineering Artificial Intelligence Networking and Parallel/Distributed Computing (SNPD), pp. 1-6, 2015.
12. https://data.gov.in/catalog/district-wise-season-wise-crop-productionstatistics

AUTHORS PROFILE

Lokesh C.K. is Assistant Professor , School of Computer Science and Applications, REVA University, India. Previously, worked as Assistant Professor and Head, Department of Computer Science, REVA Institute of Science and Managaement, Kattigenahalli, Bengaluru, India. He obtained his B.Sc. from Sri Siddaganga College of of Arts and Science,Bangalore University,Tumkur, India, in 1996, Master of Computer Applications from IGNOU University, New Delhi, India.

Dr. S. Senthil is Professor and Director, School of Computer Science and Applications, REVA University, India. Previously, he worked as Associate Professor and Head, Department of Computer Science, Vidyasagar College of Arts and Science, India. He obtained his B.Sc. (Applied Science – Computer Technology) from PSG College of Technology, India, in 1995, Master of Computer Applications from Bharathidasan University, India, in 1999, M.Phil. in Computer Science from ManonmaniamSundaranar University, India in 2002, and Ph.D. in Computer Science from BharathiarUniversity , India, in 2014. Another achievement is clearing SET in 2012. He has published 40 research papers in various reputed National and International Journals. He has presented a paper entitled "Lossless Preprocessing Algorithms for better Compression" in an IEEE International Conference at Zhangjiajie, China. His interest is in Database Systems, Data Mining, Data Compression, and Big Data Analytics.