International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 07 Issue: 07 | July 2020 www.irjet.net p-ISSN: 2395-0072

Rice Crop Yield Prediction using Recurrent Neural Networks

Chaithanya S[1], Punith Raj A[2], Rajeshrahul N[3], Mrs. Sujatha Hiremath[4], Dr. Veena Devi[5]

[1,2,3]Student, Dept. of E&C Engineering, RVCE, Karnataka, India
[4]Assistant Professor, Dept. of E&C Engineering, RVCE, Karnataka, India
[5]Associate Professor, Dept. of E&C Engineering, RVCE, Karnataka, India

---

Abstract - *Crop yield prediction is one of the most significant topics of precision agriculture. The yield of a crop depends on many different factors such as the weather, soil properties and management practices like irrigation and fertilizer usage. This paper presents a recurrent neural network (RNN) for crop yield prediction based on environmental data and historical crop yield data. The proposed RNN model, was used to forecast rice crop yield in the Karnataka state for the years 2015, 2016 and 2017. The model achieved a mean-absolute-error of 0.05187 tons per hectare for the yield prediction, which accounts for a mean absolute percentage error of 1.87796%. The model is also evaluated on the basis of a regression measure called R-squared. The model has an R-squared value of 99.9713%.*

Key Words: Crop Yield Prediction, Precision Agriculture, Deep Learning, Supervised Learning, Recurrent Neural Networks, LSTM Cells, Time Series Forecasting.

1. INTRODUCTION

Crop yield prediction is greatly important in food production and maintenance, globally. The governments can benefit from the crop yield predictions by considering the predictions made to decide upon the import and export decisions. Farmers can benefit from accurate crop yield predictions by making more informed decisions regarding crop rotations, irrigation and other management practices to produce high yields.[1] Crop yield is affected by multiple factors and it is highly challenging to make crop yield predictions. Various weather factors like the precipitation, temperature, solar insolation, surface pressure affect the crop yield very significantly. Crop yield is also dependent on soil properties and the management practices. Seed companies have significantly improved the crop genotype over the years making it even more complex to predict the crop yield.[2]

Recently, machine learning techniques like multivariate regression, decision tree, and artificial neural networks have been applied for crop yield prediction. The machine learning models consider these yield affecting factors as inputs and the crop yield as the output to identify the underlying mapping function between them, which could be a highly non-linear and complex function. Artificial neural networks are good function approximation algorithms, and can perform better in the approximation of the unknown mapping function between the crop yield and the various input variables or factors.[3]

This paper presents a recurrent neural network (RNN) for the crop yield prediction. The proposed RNN model is enhanced with long short-term memory cells (LSTM cells). The RNNs perform better for time series forecasting problems like crop yield prediction by processing the data as a series of time steps. The proposed model is trained on historical rice crop yield data of the Karnataka state along with the environmental components to make predictions for the years 2015 to 2017.

2. DATA

The data utilized for the proposed model includes yield performance and weather information. Genotype data of crops is not publicly available. The yield performance data represents the average yield of rice crop for the Karnataka state for the years 1997 to 2017. The yield data is directly available from the Government data portals. The weather data includes monthly averages of six weather factors, namely precipitation, maximum temperature, minimum temperature, surface pressure, relative humidity and all sky insolation incident on the horizontal surface. It was obtained from the NASA POWER Project which provides solar and meteorological data sets from NASA research for support of renewable energy, building energy efficiencyandagricultural needs.

3. METHODOLOGY

3.1 Data Preprocessing

The information obtained from multiple sources have to be combined into a form which is useable for the training and validation of the proposed model. The entire data was converted into multiple rows and columns, where the rows were indexed with years. For weather factors data was collected at multiple locations within the state and corresponding averages were computed in order to obtain representational values of all those factors. The weather data was available on a monthly basis, whereas the yield data was available on a yearly basis. To match the frequencies of the different data, the weather data of each month is treated as a distinct feature, making six weather factors into 72 features for the model. Finally, the complete data was representedas a

Pandas data frame with 75 unique features/columns indexed with the year.

Data scaling is an essential step in most of the machine learning models. The main aim of data scaling is to bring many different features considered for a machine learning model into a same numerical scale or range. This ensures that the machine learning model treats all the feature with same importance.[4] The standardization of the model's data was implemented by the functions from the NumPy and Pandas library. The standardization techniques have to beappliedon all the features in the data frame. This requires the calculation of the mean and standard deviation for all features individually.

Mean and standard deviation of features were calculated using NumPy functions. The Pandas data frame was hence converted into NumPy array after data visualization. The feature wise means and standard deviations were calculated from the NumPy arrays using the mean and std methods available over the NumPy arrays. These methods return NumPy arrays of mean and standard deviation feature wise. These newly obtained NumPy arrays were then utilized to make element wise operations required for the data standardization. The mean and standard deviation of all features were identified by only considering the data used while training.

Scaling of data for both training and validation data was then made from calculated means and standard deviations of the features. Each value of the data sets was subtracted from the corresponding mean and then divided by the corresponding standard deviation. This makes new data of the features to achieve 0 mean and 1 as standard deviation. The validation data set was also rescaled using the same set of means and standard deviation, because these scaleswere a part of the representations and the model also makes prediction in the same scales. The final predictions on the validation data set can be compared with its original data only when they are on the same scale. In this way the data standardization was used to rescale the data sets of the crop yield prediction models.

3.2 Yield Prediction using the Recurrent Neural Network

The crop yield prediction model is a recurrent neural network. Recurrent neural networks are a family of neural networks for processing sequential data. Recurrentnetworks share parameters in a different way. Each member of the output is a function of the previous members of theoutput.[5] Each member of the output is produced using the same update rule applied to the previous outputs. The proposed model is based on the LSTM layers which generate sequence of vectors as their outputs.

The model has a dense layer after the LSTM layers. The dense layer is a fully connected layer meaning every neuron of the dense layer is connected to every other neuron. Long Short-Term Memory networks, usually just called as LSTMs are a special kind of RNN in the previous layer. The dense layer has only one neuron and outputs a single value. The model is developed to analyze this data a time series, for better performance. LSTMs cells proved to be better performers in processing such time series data. Hence the model is designed with LSTM layers and dense layer.

The model is compiled with the loss function, optimizer function and metric for the performance. This step is significant in the model development as it governs the learning process of the model. The model compilation is followed by fitting. Fitting a model refers to the training process of the model. In the model fit step, the learning process is designed and certain callbacks are utilized for the guidance and extra functionalities during the training. All the above-mentioned design steps are experimented in multiple iterations to arrive at the optimal design specification. The model is designed to predict the yield of one future year in the yield time series, thus only one neuron is used in the dense layer. The output generated from the final dense layer is the final output of the model. This output represents the predicted yield value.

The RNN model designed for the crop yield prediction takes input data in particular dimensions, which represent the number of features in each data point, number of time steps considered and the number of data points provided for each iteration. The number of data points provided for the model at each iteration during training is called the batch size. The data is processed by the model and an output is generated. This predicted output is compared with original output using the Huber loss function, which generates the loss score. This loss score is used by the optimizer algorithm to change the weights of the model to effectively reduce the loss scores.[6] This process is iterated over all the data points for multiple times, till the optimal loss scores and hence performance is achieved.

A brief methodology of the crop yield prediction model is provided in Fig. 1. The identification of the various factors which affect the crop yield was based on the agriculture domain knowledge. The factors which can affect the yield, particularly for the rice crop are studied to identify the influential factors. The weather variables and soil properties which are influential in the crop yield were thus identified and the related data was collected. This was followed by certain data pre-processing techniques to convert it into a form usable for the neural networks. The model is designed to capture the temporal effects of all these factors. The recurrent neural network developed for this project is trained and validated using this data. The data from the year 1997 to 2011 is used for the training of the model and the data from the year 2012 to 2017 was used for the validation of the model. The mean-absolute-error was used to measure the performance of the model on the validation dataset.

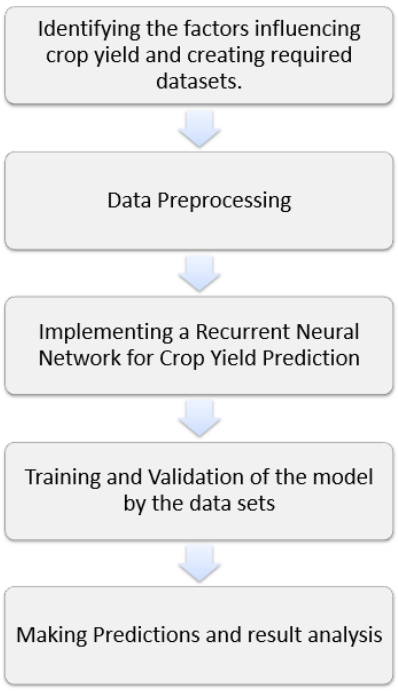

Fig -1: Design Methodology for Crop Yield Prediction

The training of the model is followed by making predictions. In this step predictions are made for the yield of rice crop for years 2015 to 2017. These predictions are then compared with the original yield values to quantify the performance of the model. This methodology encapsulates the design of the model for the rice crop yield prediction.

4. EXPERIMENTATION WITH LEARNING RATE

The learning rate of the optimization algorithm is the metric by which the algorithm changes the weights of the neural network. The learning rate affects the quality of training and the effective time taken for the training.[7] Hence, this hyper parameter is of high importance in training a neural network. If the learning rate is too low, the weights of the neural network are given tiny updates and the learning process will progress very slowly. This will more time to train the neural network to reach optimal performance. The other way, if the learning rate is too high, the weights are updated drastically causing the model to behave undesirably. Hence, finding an optimal learning rate improves the performance of the model and also speeds up the training process. The optimal learning rate for the training depends on the model architecture and data usedin the training. This experimentation is based on the idea of exposing the model for a range of learning rates during the training. At the training phase the model is trained at varying learning rates and the loss scores are recorded at all those learning rates. This information is then used to identify the optimal learning rate.

The learning rate is initially set to a very low value. The very low learning rates are not optimal for the training of the neural network as they proceed very slowly and hence do not work with large data sets. After starting with a very low learning rate, at each training iteration, the learning rate is increased. The loss scores are recorded at each iteration when the learning rates are changed. The learning rate is increased until it reaches a very high value, whichpractically cannot obtain any acceptable performance for the model. The loss scores at such high learning rates explode to very high values.

Typically, the low and high upper bounds for the learning rate are 1e-10 and 1e+1. The learning rate varying procedure results in a useful information to identify the optimal learning rate. The learning rate and the loss scores recorded at each of those learning rates are plotted. The plot is made for the entire range learning rate used in the training of the model. The graph obtained from this information has certain characteristics in terms of its shape and in particular its varying slopes. The graph obtained from the experimentation made on the crop yield prediction is provided in the Fig. 2.

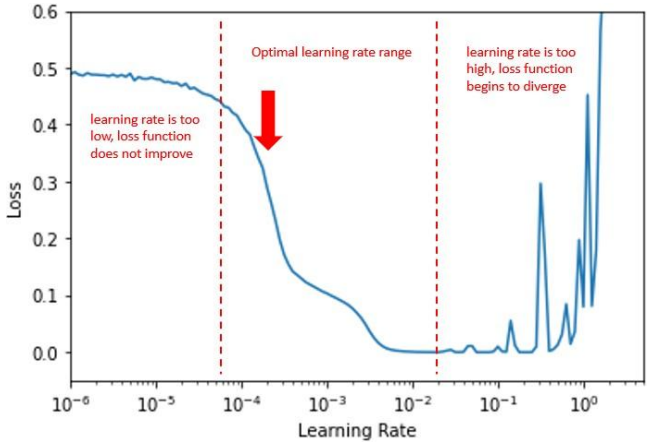

Fig -2: Finding an Optimal Learning Rate

The optimal learning rate was identified from the steep drop observed in the plot. The model was trained by a learning rate of 2e-4. This value of learning rate was optimal for this model. The training of this model obtained best performance on validation data set in less amount of training.

5. RESULTS AND ANALYSIS

The performance of the crop yield prediction model is quantified by three different metrics. The metrics used for the performance measure of the crop yield prediction model are mean absolute error (MAE), mean absolute percentage error (MAPE) and R-squared. The mean absolute error was calculated on the three predictions made by the model and the actual yield values for the years 2015, 2016 and 2017. It measures the average of the magnitudes of the errors in the

forecasts made by the model. The mean absolute error measure the average model prediction error in the units comparable with the actual predictions made. This does not consider the direction of the errors. The mean absolute error is more widely used for expressing prediction accuracies. The model exhibits a mean absolute error of 0.05187 in units of tons per hectare. The mean absolute error of the crop yield prediction model was calculated by using the corresponding metric function of the Keras library. The other metric used over the crop yield prediction model is the mean absolute percentage error. The mean absolute percentage error is similar to the mean absolute error in a sense that it does not consider the direction of the error. The mean absolute percentage error gives the measure of the error of the model prediction relative to the actual values. This is more helpful in comparing the results of the model with another yield prediction model. The mean absolute percentage error of the crop yield prediction is 1.87796. It can be interpreted as 1.8779%. The mean absolute percentage error essentially calculates the mean of the percentage errors in the predictions made by the model.

The mean absolute percentage error of the crop yield prediction model was calculated by using corresponding metric function from the Keras library. The performance of the model was also expressed by calculating the R-squared measure for the model predictions. The R-squared value is a statistical measure which calculates the similarity the regression line and the data it is fitted to. The values of the R-squared is always between 0 and 1. If the R-squared value is 1, then the model 100% predicts the data variance and if the R-squared value is 0, then the model predicts none of the variance.[8] In statistical definitions R-squared is defined as the ration of the explained variance of the model to the total variance of the target variable. Essentially R-squared measures the strength of the relationship between the crop yield prediction model and the actual yield values on a 0 to 1 scale. The R-squared measure was calculated to be 0.999713. It can be interpreted as 99.9713 percentage.

Table -1: Performance Metrics

Metric	Value
Mean Absolute Error	0.05187 Tons/Hectare
Mean Absolute Percentage Error	1.87796 %
R-Squared	99.9713 %

The values of all the performance metrics considered for the crop yield prediction model are provided in Table 1. The R-squared measure explains how the predicted values of yield are scattered with respect to the regression line. For the better understanding of the R-squared measure and more importantly the prediction performance of the crop yield prediction model, a plot was plotted between the actual yield values and the predicted yield values. The plot explains how the predicted values scatter around the regression line. This plot was generated using SciPy library functions along with Matplotlib library functions. The function generated slope and intercept for the regression line based on the actual and predicted yield values. These slope and intercept values were used with the plot function of the Matplotlib library to generate the required regression line. The scatter plot of the actual yield values and the predicted yield values were fitted on top of the regression line. These plots can explain the prediction performance of the crop yield prediction model.

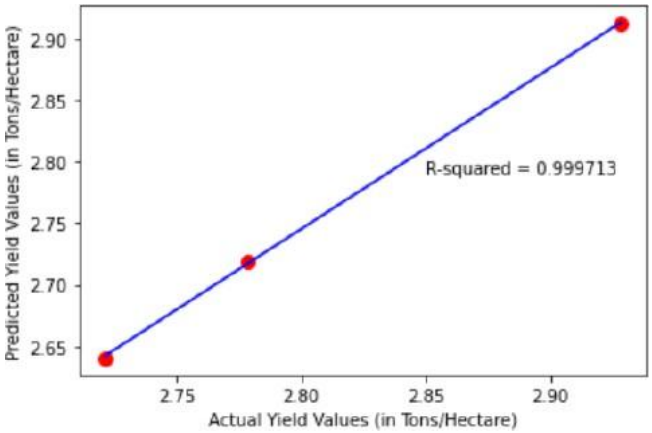

Fig -3: Actual vs Predicted Yields

These plots in the Fig. 3 depicts the scatter plot between the actual yield values and the values predicted by the crop yield prediction model. The x-axis denotes the actual yield values and the y-axis denotes the predicted yield values. The red dots are the represents the scatter plot between the actual and predicted yield values while the blue line is the regression line. It can be noted from the plot that the scatter plot points are not so far from the regression line, which represents a better prediction performance of the crop yield prediction model. The values of the R-squared is also mentioned with the plot.

6. CONCLUSIONS

The crop yield prediction model was successfully implemented for rice crop yield predictions of the Karnataka state. Data for the crop yield prediction model included historical information of rice yield performance along with various other environmental components data of Karnataka. The recurrent neural network was implemented in python by using TensorFlow and Keras libraries. The training of the model was done after finding an optimal learning rate by certain experimentation which was identified to be 2e-4. The yield of rice crop was predicted for the years 2015, 2016 and 2017. The crop yield prediction model exhibits a mean absolute error of 0.05187 tons per hectare. The mean absolute percentage error of the model is 1.87796%. The model is also evaluated on the basis of a regression measure called R-squared. The R-squared value of the model is 99.9713%. The crop yield prediction model can be used for

predicting the rice crop yield of Karnataka for the future years. The model can make a yield prediction for a year, if it is given the required data for three previous years.

REFERENCES

[1] S. Khaki and L. Wang, "Crop yield prediction using deep neural networks", Frontiers in Plant Science, vol. 10, May 2019.

[2] S. Khaki, L. Wang, and S. V. Archontoulis, "A cnn-rnn framework for crop yield prediction", Frontiers in Plant Science, vol. 10, Jan. 2020.

[3] C. Jiang D., X. Yang, N. Clinton, and N. Wang, "An artificial neural network model for estimating crop yields using remotely sensed information", International Journal of Remote Sensing, vol. 25, no. 9, pp. 1723-1732, May 2004.

[4] S. T. Drummond, K. A. Sudduth, A. Joshi, S. J. Birrell, and N. R. Kitchen, "Statistical and neural methods for site-specific yield prediction", Transactions of the ASAE, vol. 46, no. 1, p. 5, 2003.

[5] X. Glorot and Y. Bengio, "Understanding the difficulty of training deep feedforward neural networks", in Proceedings of the thirteenth international conference on artificial intelligence and statistics, 2010, pp. 249-256.

[6] J. Liu, C. Goering, and L. Tian, "A neural network for setting target corn yields", Transactions of the ASAE, vol. 44, no. 3, p. 705, 2001.

[7] I. Goodfellow, Y. Bengio, and A. Courville, Deep learning. MIT press, 2016.

[8] K. Suresh and S. Krishna Priya, "A study on pre-harvest forecast of sugarcane yield using climatic variables", Stat. Appl. 7&8 (1&2) (New Series), 2009.