International Journal of Engineering Technology and Management Sciences
Website: ijetms.in Issue: 4 Volume No.6 Aug-Sept – 2022
DOI:10.46647/ijetms.2022.v06i05.043 ISSN: 2581-4621

CROP YIELD PREDICTION USING DEEP XGBOOST ALGORITHM

Ms. S. ABARNA[1] , P. GANESH PRIYA[2]
[1]Assistant Professor, Dept. of Computer Science and Engineering, Grace College of Engineering, Tuticorin
[2]PG Student, Dept. of Computer Science and Engineering, Grace College of Engineering, Tuticorin

ABSTRACT: Predicting crop yield is a complex task since it depends on multiple factors. Although many models have been developed so far in the literature, the performance of current models is not satisfactory, and hence, they must be improved. In this study, we developed deep learning-based models to evaluate how the underlying algorithms perform with respect to different performance criteria. The algorithms evaluated in our study are the XGBoost machine learning (ML) algorithm, Convolutional Neural Networks (CNN),XGBoost, and Recurrent Neural Networks (RNN). For the case study, we predicted crop yield based on the environmental, soil, silt, nitrogen, clay, ocd, ocs, pHH2O, sand, soc, ceo, water and crop parameters has been a potential research topic. Our proposed method has high performance by preforming feature selection on predicting the crop yield. Our proposed method used for both the soyabeans and corn crop for predicting their yield.
Keywords:XGBOOST,Cropyield Learning algorithms

INTRODUCTION

Crop yield prediction is helpful for farmers, Capture the time dependencies of environmental factors and the genetic improvement of seeds.Yield prediction for untested environments without significant drop in the prediction accuracy.

1.1 X G B O O S T

XGBoost is a decision-tree-based ensemble Machine Learning algorithm that uses a gradient boosting framework. In prediction problems involving unstructured data (images, text, etc.) artificial neural networks tend to outperform all other algorithms or frameworks. However, when it comes to small-to-medium structured/tabular data, decision tree based algorithms are considered best-in-class right now. Please see the chart below for the evolution of tree-based algorithms over the years. As a result, there is a strong community of data scientists contributing to the XGBoost open source projects with ~350 contributors and ~3,600 commits on GitHub. XGBoost stands for Extreme Gradient Boosting. It uses more accurate approximations to find the best tree model.

Boosting: N new training data sets are formed by random sampling with replacement from the original dataset, during which some observations may be repeated in each new training data set

The algorithm differentiates itself in the following ways:

A wide range of applications: Can be used to solve regression, classification, ranking, and user-defined prediction problems.

Portability: Runs smoothly on Windows, Linux, and OS X.

Languages: Supports all major programming languages including C++, Python, R, Java,

Cloud Integration: Supports AWS, Azure, and Yarn clusters and works well with Flink, Spark, and other ecosystems.

Regularization: It penalizes more complex models through both LASSO (L1) and Ridge (L2) regularization to prevent over fitting.

Sparsity Awareness: XG Boost naturally admits sparse features for inputs by automatically 'learning' best missing value

International Journal of Engineering Technology and Management Sciences
Website: ijetms.in Issue: 4 Volume No.6 Aug-Sept – 2022
DOI:10.46647/ijetms.2022.v06i05.043 ISSN: 2581-4621

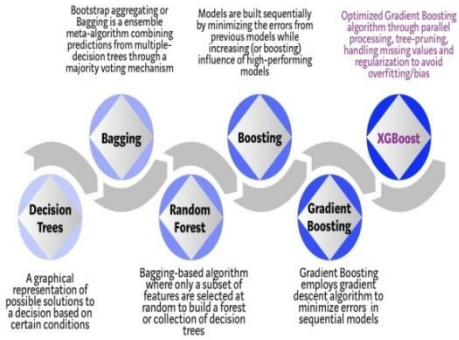

Fig1.1 Algorithmic Enhancements

Supervised learning

XGBoost is used for supervised learning problems, where we use the training data (with multiple features) to predict a target variable. Before we learn about trees specifically, let us start by reviewing the basic elements in supervised learning.

XGBoost is a set of open source functions and steps, referred to as a library, that use supervised ML where analysts specify an outcome to be estimated/ predicted. The XGBoost library uses multiple decision trees to predict an outcome. The ML system is trained using batch learning and generalized through a model based approach.

It uses all available data to construct a model that specifies the relationship between the predictor and outcome variables, which are then generalized to the test data.

Decision trees

The downside of using XGBoost compared to a neural network, is that a neural network can be trained partially whereas an XGBoost regression model will have to be trained from scratch for every update. This is because an XGBoost model uses sequential trees fitted on the residuals of the previous trees so iterative updates to the model are not really possible.

Important Parameters of XGBoost

Booster: (default=gbtree)

It is based one the type of problem (Regression or Classification)

Gbtree/dart – Classification , gblinear – Regression.

N threads: (default – it is set maximum number of threads available)

Number of parallel threads needed to run XGBoost.

Early stopping_ rounds :

XGBoost supports early stopping after a fixed number of iterations.

In addition to specifying a metric and test dataset for evaluation each epoch, you must specify a window of the number of epochs over which no improvement is observed. This is specified in the early stopping rounds parameter.

Eta η (default = 0.3, alias : learning rate, range : [0,1])

Step size shrinkage used in update to prevents overfitting. After each boosting step, we can directly get the weights of new features, and eta shrinks the feature weights to make the boosting process more conservative.

Gamma (default=0, alias: min_split_loss, range: [0,∞])

Minimum loss reduction required to make a further partition on a leaf node of the tree. The larger gamma is, the more conservative the algorithm.

Working of XGBoost

In order to understand XGBoost, we must first understand Gradient Descent and Gradient Boosting.

a) Gradient Descent:

A cost function measures how close the predicted values are, to the corresponding actual values.

International Journal of Engineering Technology and Management Sciences
Website: ijetms.in Issue: 4 Volume No.6 Aug-Sept – 2022
DOI:10.46647/ijetms.2022.v06i05.043 ISSN: 2581-4621

Ideally, we want as little difference as possible between the predicted values and the actual values. Thus, we want the cost function to be minimized.The weights associated with a trained model, cause it to predict values that are close to the actual values. Thus, the better the weights associated with the model, the more accurate are the predicted values and the lower is the cost function. With more records in the training set, the weights are learned and then updated. Gradient Descent is an iterative optimization algorithm. It is a method to minimize a function having several variables. Thus, Gradient Descent can be used to minimize the cost function. It first runs the model with initial weights, then seeks to minimize the cost function by updating the weights over several iterations.

b) Gradient Boosting:

Boosting: An ensemble of weak learners is built, where the misclassified records are given greater weight ('boosted') to correctly predict them in later models. These weak learners are later combined to produce a single strong learner. There are many Boosting algorithms such as AdaBoost, Gradient Boosting and XGBoost. The latter two are tree-based models. Figure 1 depicts a Tree Ensemble Model.

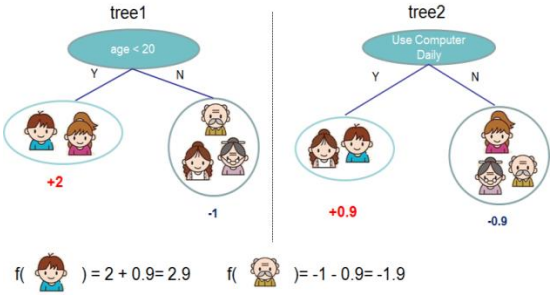

Figure 1.2: Tree Ensemble Model to predict whether a given user likes computer games or not. +2, +0.1, -1, +0.9, -0.9 are the prediction scores in each leaf.

The final prediction for a given user is the sum of predictions from each tree. Source Gradient Boosting carries the principle of Gradient Descent and Boosting to supervised learning. Gradient Boosted Models (GBM's) are trees built sequentially, in series. In GBM's, we take the weighted sum of multiple models.

Each new model uses Gradient Descent optimization to update/ make corrections to the weights to be learned by the model to reach a local minima of the cost function.

The vector of weights assigned to each model is not derived from the misclassifications of the previous model and the resulting increased weights assigned to misclassifications, but is derived from the weights optimized by Gradient Descent to minimize the cost function. The result of Gradient Descent is the same function of the model as the beginning, just with better parameters.

Gradient **Boosting** adds a new function to the existing function in each step to predict the output. The result of Gradient Boosting is an altogether different function from the beginning, because the result is the addition of multiple functions.

c) XGBoost:

XGBoost was built to push the limit of computational resources for boosted trees. XGBoost is an implementation of GBM, with major improvements. GBM's build trees sequentially, but XGBoost is parallelized. This makes XGBoost faster.

Features of XGBoost:

XGBoost is scalable in distributed as well as memory-limited settings. This scalability is due to several algorithmic optimizations.

1. Split finding algorithms: approximate algorithm:

To find the best split over a continuous feature, data needs to be sorted and fit entirely into memory. This may be a problem in case of large datasets.

An approximate algorithm is used for this. Candidate split points are proposed based on the

DOI:10.46647/ijetms.2022.v06i05.043 ISSN: 2581-4621

percentiles of feature distribution. The continuous features are binned into buckets that are split based on the candidate split points. The best solution for candidate split points is chosen from the aggregated statistics on the buckets.

2. Column block for parallel learning:

Sorting the data is the most time-consuming aspect of tree learning. To reduce sorting costs, data is stored in in-memory units called 'blocks'. Each block has data columns sorted by the corresponding feature value. This computation needs to be done only once before training and can be reused later. Sorting of blocks can be done independently and can be divided between parallel threads of the CPU. The split finding can be parallelized as the collection of statistics for each column is done in parallel.

3. Weighted quantile sketch for approximate tree learning:

To propose candidate split points among weighted datasets, the Weighted Quantile Sketch algorithm is used. It carries out merge and prune operations on quantile summaries over the data.

4. Sparsity-aware algorithm:

Input may be sparse due to reasons such as one-hot encoding, missing values and zero entries. XGBoost is aware of the sparsity pattern in the data and visits only the default direction (non-missing entries) in each node.

5. Cache-aware access:

To prevent cache miss during split finding and ensure parallelization, choose 2^16 examples per block.

6. Out-of-core computation:

For data that does not fit into main memory, divide the data into multiple blocks, and store each block on the disk. Compress each block by columns and decompress on the fly by an independent thread while disk reading.

7. Regularized Learning Objective:

To measure the performance of a model given a certain set of parameters, we need to define an objective function. An objective function must always contain two parts: training loss and regularization. The regularization term penalizes the complexity of the model.

$Obj(\Theta)=L(\theta)+ \Omega(\Theta)$

where Ω is the regularization term which most algorithms forget to include in the objective function. However, XG Boost includes regularization, thus controlling the complexity of the model and preventing overfitting.

MODULES

Crop yield prediction using deep reinforcement learning method can be implemented by using the following modules. In each estimation is help to improve the crop yield prediction.

- Convolutional Neural Network
- Recurrent Neural Network
- XG Boost

DESIGNARCHITECTURE

In design architecture the following process are required as Fig(5.1)

FEATURE SELECTION

In machine learning and statistics, feature selection, also known as variable selection, attribute selection or variable subset selection, is the process of selecting a subset of relevant features (variables, predictors) for use in model construction. Feature selection techniques are used for several reasons: simplification of models to make them easier to interpret by researchers/users,

shorter training times,

To avoid the curse of dimensionality, improve data's compatibility with a learning model class

International Journal of Engineering Technology and Management Sciences
Website: ijetms.in Issue: 4 Volume No.6 Aug-Sept – 2022
DOI:10.46647/ijetms.2022.v06i05.043 ISSN: 2581-4621

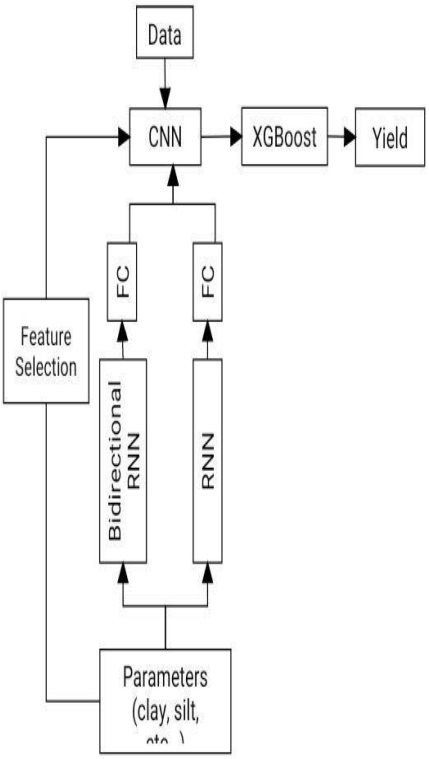

Fig.5.1SystemDesign Architecture

Encode inherent symmetries present in the input space.
The central premise when using a feature selection technique is that the data contains some features that are either redundant or irrelevant, and can thus moved without incurring much loss of information. Redundant and irrelevant are two distinct notions, since one relevant feature may be redundant in the presence of another relevant feature with which it is strongly correlated. Feature selection techniques should be distinguished from feature extraction. Feature extraction creates new features from functions of the original features, whereas feature selection returns a subset of the features. Feature selection techniques are often used in domains where there are many features and comparatively few samples (or data points). Archetypal cases for the application of feature selection include the analysis of written texts and DNA microarray data, where there are many thousands of features, and a few tens to hundreds of samples.

CNN

CNN- Convolutional Neural Network

In deep learning, a convolutional neural network is a class of deep neural network used in image recognition and processing that is specifically designed to process pixel data .CNNs are used toanalysethevisualimage.Ithas6 layers.

It contains,

- Convolution layer
- Batch normalization layer
- Max pooling layer,
- Dropout layer
- Flatten layer.

International Journal of Engineering Technology and Management Sciences
Website: ijetms.in Issue: 4 Volume No.6 Aug-Sept – 2022
DOI:10.46647/ijetms.2022.v06i05.043 ISSN: 2581-4621

- Fully connected
- Convolution layer

The first layer of a Convolutional Neural Network is a always a Convolutional Layer. This layer is applies a filter only to the input image to extracts the features from the input image with the filter of particular size M*M. By sliding the filter over the input image, the dot product is taken between the filter and to part so the input image with respect to the filter.
It gives the output as corners and edges of the input image.

Batch normalization layer

Batch normalization typically behaves differently in training ode and prediction mode. It resizes or rescales and re-center the image.

Max pooling layer

The largest element is taken from feature map. This layer reduces the spatial size of the features and reduces over fitting and provide abstract representation. The pooling layerusuallyservesasabridgebetweentheconvolutionallayerandtheFClayer

Dropout layer

Dropout is only used after the pooling, but this is just rough heuristic. This layer prevents the model form over fitting .The dropout layers randomly sets input units to 0 with a frequency of rate at each step during training.
The drop out layer only applies when training is set to true such that no values are dropped during inference.
Flattening is converting the data into a1-dimensional array for inputting it to the next layer

Fully connected layer

The fully connected layer consists of the weights and biases along with the neurons between two different layers. The input image from the previous layers are flattened and fed to the FC layer.

CNN Advantage

It automatically detects the important features without any human supervision.

RNN

RNNs are known as Recurrent Neural Network. RNN are a type of neural network where the output from the previous step is fed as input to the current step. RNN is a type of artificial neural network which uses sequential
Data or time series data. It has2 layers.
It contains
i. Flatten layer
ii. Dense layer.
Flatten layer
This layer collapses the spatial dimensions of the input into channel dimension.

Dense layer

Dense layer is the regularly deeply connected neural network layer. Dense layer does below operation on the input and return the output . This layer collects all the output from the previous layer.

RNN Advantage

RNN can process inputs of any length
Even if the input size is larger, them Resize does not increase.

XG BOOST

XG Boost or extreme gradient boosting is one of the well-known gradient boosting techniques(ensemble) having enhanced performance and speed in tree-based (sequential decision trees) machine learning algorithms.

DOI:10.46647/ijetms.2022.v06i05.043 ISSN: 2581-4621

XGBoost was created by Tianqi Chen and initially maintained by the Distributed (Deep) Machine Learning Community (DMLC) group.
It is the most common algorithm used for applied machine learning in competitions
Has gained popularity through winning solutions in structured and tabular data.

CONCLUSION

Presented a machine learning approach for crop yield prediction. The approach used deep neural networks to make yield predictions based on environment data. The carefully designed deep neural networks were able to learn nonlinear and complex relationships between genes, environmental conditions, as well as their interactions from historical data and make reasonably accurate predictions of yields for new hybrids planted in new locations with known weather conditions. Performance of the model was found to be relatively sensitive to the quality of weather prediction, which suggested the importance of weather prediction techniques. A major limitation of the proposed model is its black box property, which is shared by many machine learning methods. The feature selection approach successfully found important features, and revealed that environmental factors had a greater effect on the crop yield than genotype.
Whenever time and technology changes everything needs to be changed as enhanced. In future the system can be expanded in the following services. The system is developed in such away that any further modification to system can be achieved without any great alteration in program structure. Our future research is to overcome this limitation by looking form readvanced models that are not only.

FUTUREENHANCEMENT

Whenever time and technology changes everything needs to be changed as enhanced. In future the system can be expanded in the following services. The system is developed in such a way that any further modification to system can be achieved without any great alteration in program structure. Our future research is to overcome this limitation by looking for more advanced models that are not only more accurate but also more explainable. In future the work is to predict and extract the crop yield with an extended algorithm for more advanced and intelligent way of learning and predicting the results. This system can further extend to finding the yields with less discriminative data and high prediction power

Reference

Abbaszadeh,P.,Gavahi, K., Alipour, A., Deb, P. and Moradkhani, H., 2022. Bayesian multi-modeling of deep neural nets for probabilistic crop yield prediction. Agricultural and Forest Meteorology, 314, p.108773.
Ali, A.M., Abouelghar, M.A., Belal, A.A., Saleh, N., Younes, M., Selim, A., Emam, M.E., Elwesemy, A., Kucher, D.E. and Magignan, S., 2022. Crop Yield Prediction Using Multi Sensors Remote Sensing. The Egyptian Journal of Remote Sensing and Space Science.
Fan, J., Bai, J., Li, Z., Ortiz-Bobea, A. and Gomes, C.P., 2022, June. A GNN-RNN approach for harnessing geospatial and temporal information: application to crop yield prediction. In Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 36, No. 11, pp. 11873-11881).
Gupta, S., Geetha, A., Sankaran, K.S., Zamani, A.S., Ritonga, M., Raj, R., Ray, S. and Mohammed, H.S., 2022. Machine Learning-and Feature Selection-Enabled Framework for Accurate Crop Yield Prediction. Journal of Food Quality, 2022.
Moswa, A., 2022. Corn Yield Prediction Using Crop Growth and Machine Learning Models (Doctoral dissertation, Université d'Ottawa/University of Ottawa).
Oikonomidis, A., Catal, C. and Kassahun, A., 2022. Hybrid Deep Learning-based Models for Crop Yield Prediction. Applied artificial intelligence, pp.1-18.
Shuai, G. and Basso, B., 2022. Subfield maize yield prediction improves when in-season crop water

DOI:10.46647/ijetms.2022.v06i05.043 ISSN: 2581-4621

deficit is included in remote sensing imagery-based models. Remote Sensing of Environment, 272, p.112938.
Ziliani, M.G., Altaf, M.U., Aragon, B., Houborg, R., Franz, T.E., Lu, Y., Sheffield, J., Hoteit, I. and McCabe, M.F., 2022. Early season prediction of within-field crop yield variability by assimilating CubeSat data into a crop model. Agricultural and Forest Meteorology, 313, p.108736.