IJNRD.ORG ISSN : 2456-4184
INTERNATIONAL JOURNAL OF NOVEL RESEARCH AND DEVELOPMENT (IJNRD) | IJNRD.ORG
An International Open Access, Peer-reviewed, Refereed Journal

A SURVEY FROM MACHINE LEARNING TO DEEP LEARING ALGORITHMS ON AGRICULTURE

[1]Mrs. E.KANIMOZHI M.Sc., MPhil., [2]Dr. T. PRABHU MCA., MPhil., ME., PhD

[1]Research Scholar, Department of Computer Science, Dr. MGR Educational & Research Institute, Maduravoyal, Chennai - 600095.

[2]Associate Professor & Deputy HOD, Department of Computer Applications, Dr. MGR Educational & Research Institute, Maduravoyal, Chennai - 600095.

Abstract

Crop yield has been forecasted using ecological, soil, water, and crop nutrient proportions in a probable research design. A variety of factors such as weather circumstances, earth quality, irrigation and level of ground water should be considered while forecasting the crop yield. Numerous models based on deep learning are used to mine handy information about the crop yield. The proposed model implements an Ensemble based Deep Learning model for forecasting the crop yield based on the crop nutrient proportion and other external factors. The result comparison ensures that the proposed model is better in forecasting the crop yield with higher accuracy and minimal errors than various existing models.

Introduction

The crop yield depends on factors like crop genotype, surroundings, the methodologies used in agriculture process and weather conditions. The crop genotype has enhanced considerably over the recent years by various seed companies where as the other factors such as surroundings and weather conditions affect the crop yield to an extent [1]. The outcome of the interaction among the surrounding and genotype in a detailed environmental study resulted in dividing surroundings into similar groups [2]. A precise focus on water administration and supply models overseen by random forest methodology found out to enhance the forecasting in mongo yield [3].

In ML models various stages such as preprocessing, feature extraction, feature selection and classification are used. Any noise that is included as a part of the image can be removed by the process of preprocessing. The preprocessed image is fed to the feature extraction module to extract the possible features from the data and the final classifier selects the most vital features from the set of features. The final decision is completed by means of a separate classifier in the classification stage. In DL based models, the feature extraction and classification stages are fused as a single stage in the deep model as represented in Figure 1.1.

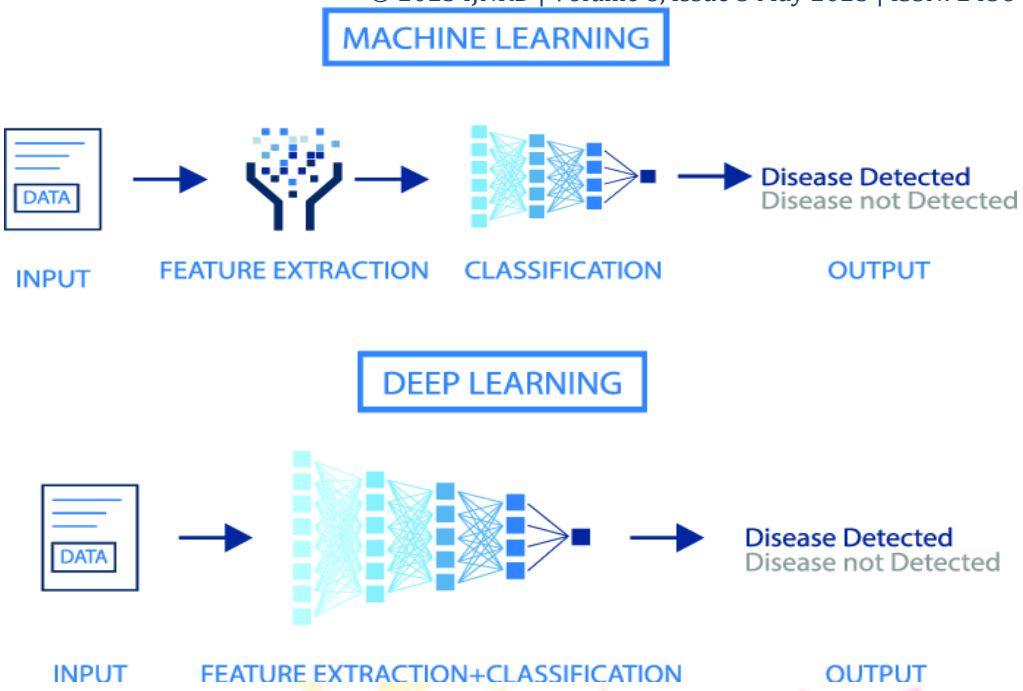


Fig. 1.1: Architecture of ML and DL models

Precise forecasting on the crop yield can help in planning the import and export operations and the same can be used by the farmers to make financial decisions. The crops with hybrid genotype are likely to produce more yield new surroundings. It is also noted that forecasting of the crop yield is not 100% accurate as it depends on multiple factors rather than an individual one. Random forest and manifold linear regression algorithms were implemented to forecast the yield of crops such as Wheat, Maize and Potato. Artificial Neural Network (ANN) was utilized to imitate the relation between the dependence of maize yield and the surroundings [4]. A numerical optimization technique and multiple linear regression technique were used to forecast the potato and wheat yield using biomass calculation [5]. Numerous techniques and approaches are used in forecasting the crop yield prediction by employing Deep Learning (DL) models to improve the methodologies used in agriculture process thereby enhancing the crop yield. The paper is organized as follows: Section II provides a study about various existing models for crop yield prediction. The proposed model is discussed in Section III. The next section provides the results and discussion. The paper is concluded with future research prospective.

Existing Works:

More recently, machine learning techniques have been applied for crop yield forecasting, including multivariate regression, decision tree, association rule mining, and ANN. A salient feature of machine learning models is that they treat the output (crop yield) as an implicit function of the input variables (genes and environmental components), which could be a highly non-linear and complex function. A weighted histograms regression to forecast the yield of different soybean varieties, which demonstrated superior performances over conventional regression algorithm, was implemented [6].

Compared with the aforementioned neural network models in the literature, which were shallow networks with a single hidden layer, deep neural networks with multiple hidden layers are more powerful to reveal the fundamental non-linear relationship between input and response variables [7]. Bayesian Neural Network (BNN) is utilized to predict future climate. Late-season forecasting in the US Corn Belt using a BNN model has an R2 of 0.77, above the average coefficient of determination (R2) in testing years 2010 to 2019. Under both normal and exceptional weather circumstances, the suggested BNN model accurately predicted the production of maize [8].

DL is a subset of ML that is analogous to the human brain in processing the information. Rather than using the rule based approach for the operation, DL maps the data to the labels and processes them. DL

algorithms are comprised of multiple layers which provide different dimensions of data as it passes through them as suggested [9].

Ensemble Learning Models

A generic ensemble learning model is made up of numerous steps with every task consisting of a dataset, an inducer and a classifier [10]. An ensemble prediction model consists of three steps that involve constructing a diverse set of base models, followed by a voluntary step to prune the models using heuristics. The final step is creating the final output by prediction combination. These three steps can be modeled as a problem of optimization and a solution can be arrived. Generating members of ensemble model [11] in predicting time series data and for the classification of imbalanced data and identifying fault [12].

Support Vector Machine:

Of the different pattern recognition areas, Support Vector Machine (SVM) is reliably in use to minimize the structural risk. In the decision surface, i.e. in the hyper plane support vectors reveal the nearness of data points. The purpose of SVM is to create a model keeping training data as the basis. It yields determined values based on the test data attributes [13].

Naive Bayes:

Based on Bayesian theorem the Naive Bayes classification technique is developed. When the value of inputs is very high, this technique is most suitable. Simple Bayes or Idiot Bayes are the other names of Bayes classifiers [10].

Decision Tree Regressor:

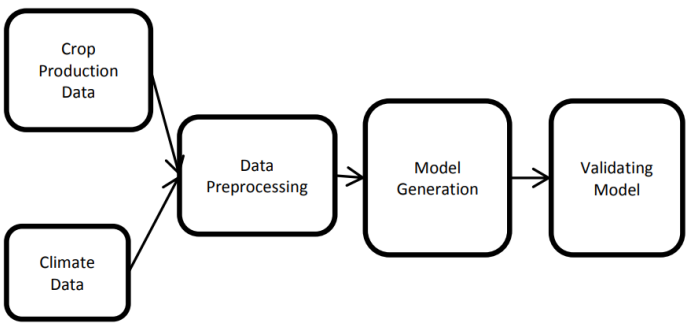
Decision tree commonly starts with a single node, which further splits into many possible nodes. Every node develops other additional nodes, which further split and lead into other possibilities. The root node is item crop under consideration, which is the most dominant feature in the model.

Working of AdaBoost Regressor:

The idea of using boosting algorithm is to improve the weak learner to generate a better result. An algorithm whose performance is very weak to get learned is known as weak learner. Training many individual models in a sequential way is known as boosting. Every model learns from errors made by the former model. The drawback of AdaBoost is decision trees with one split. AdaBoost calculates the weight of observations in such a way that it assigns more weight on instances which are hard to classify and less weight on instances which are already handled.

Proposed Ensemble Learning Model

The overview of the proposed model is represented in Figure 1.2.



Crop yield prediction is a system which is used to predict the suitable crop based on the parameters considered in the dataset. Ensembling of algorithms improves the accuracy rate rapidly. The process consists of the following modules:

Pre-processing:

In the dataset there are two categorical columns. Categorical data are the attributes that where the values are in labelled format. There is a fixed set which doesn't exceed the limit of the possible values, here in this case, crop name and country values. Multiple machine learning algorithms cannot operate on labelled data directly. In order to operate the label, data should be converted to numerical data using any one of the encoding techniques.

Data Exploration:

This is done by exploring the relationships between the columns of the dataset. A best way to check correlations among columns is by visualizing the correlation matrix using a heat map. Pearson Correlation is used to visualize the correlation matrix as a heat map.

Scaling:

The attributes in the data frame are highly different in range, magnitudes, and units. The features with high values will result in huge complexity and confusion rather than features with smaller values. To overcome this, we need to bring all features to same range of magnitudes. This can be done by scaling the data with MinMaxScaler or StandardScaler.

Training and Testing Phase:

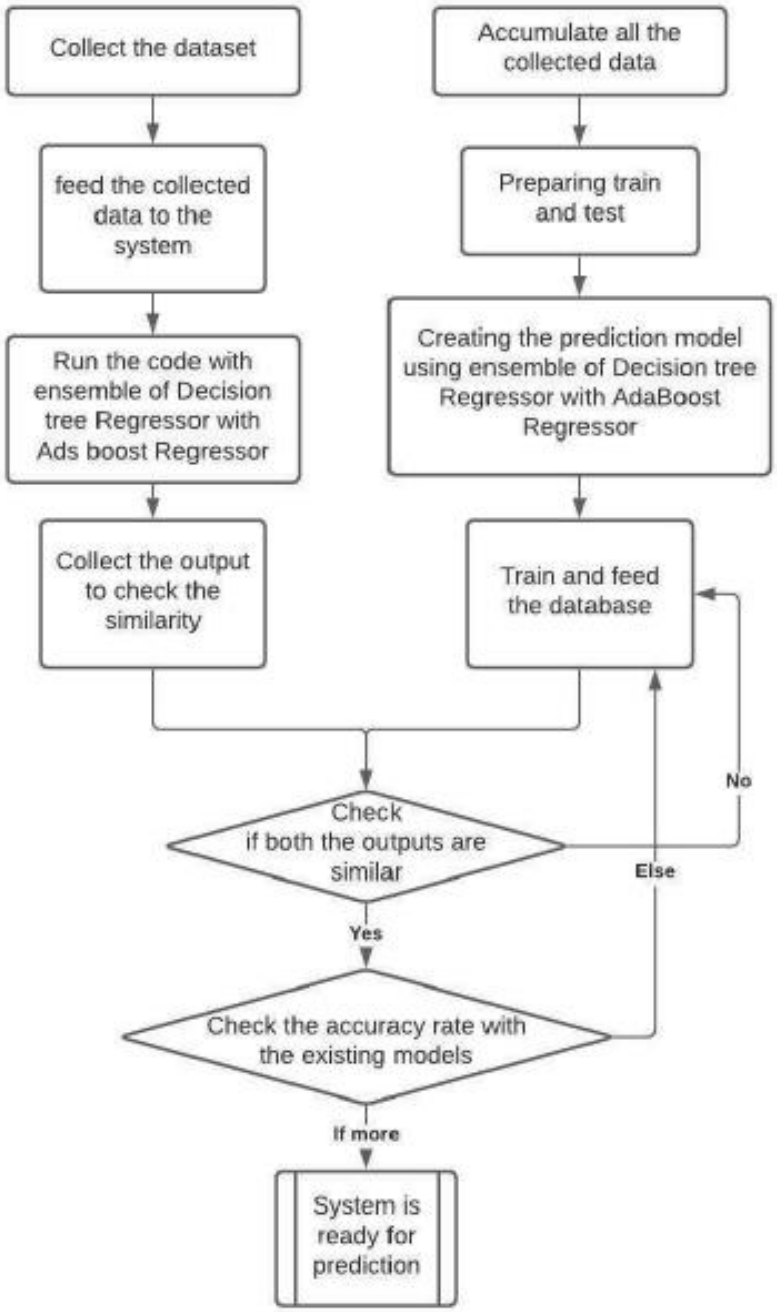
The crucial step on data preprocessing is the training and testing data. Here, we have considered 70/30 ratio which is the most considered ratio. This predicts how well the test data is trained giving the most accurate outcome.

Model Comparison & Selection:

Before deciding to choose an algorithm to use, evaluation should be done to choose the best one that fits for the specific dataset. Basically, when working on a machine learning problem with a given dataset, we try various models to solve optimization problem. But the suitable model will neither over fit nor under fit the model. Regression Analysis is a form of prediction technique which determines the relationship between a dependent and independent variable.

Here, we will compare few Regression and Classification models through their Rooted Square Value:

- Decision Tree Regressor
- Random Forest Regressor
- Gradient Boosting Regressor
- AdaBoost Regressor
- K- nearest Neighbor Classifier
- Bagging Classifier



Two most used ensemble methods are boosting and bagging:

1) Boosting: This is a technique where every individual model is trained in a sequential way, where every model generated will learn from the mistakes of formerly produced model.

2) Bagging: This is a technique where every individual model is trained in a sequential way, where every model generated from subset data.

Ensembling Decision Tree Regressor with Random Forest Regressor:

These two algorithms have a similar parameter called maximum depth. So, based on the maximum depth, we should combine these two algorithms to get better results. The accuracy is 95% for decision tree when the maximum depth is 25. So, the same depth is given as input parameter for the Random Forest. R2 Score value is evaluated for yield actual and yield predicted.

Ensembling Decision Tree Regressor with Ada Boost Regressor:

Boosting: This is a technique where every individual model is trained in a sequential way, where every model generated will learn from the mistakes of formerly produced model.

AdaBoost - is a boosting ensembling model which is robust while working with decision trees. Boosting model's technique is grasping from the prior mistakes.

Ensembling Bagging Classifier with K-Nearest Neighbor:

Bagging: This is a technique where every individual model is trained in a sequential way, where every model generated from subset data. This is done by combining these two models by giving KNN as a base estimator for the Bagging Classifier.

Ensembling Random Forest Regressor with Ada Boost Regressor:

Random Forest Regressor using Ada Boost results in the range of "0.81 to 0.94", AdaBoost is boosting algorithm that binds itself unto other functional interplays. Using random forest as parameter for AdaBoost gives the better result with 94% accuracy compared to the individual prediction.

Ensembling Decision Tree Regressor with Gradient Boosting Regressor:

Gradient boosting is one of the most strong and robust techniques for developing forecasting models. Performance of decision tree is increased by boosting up the algorithm with gradient boosting. This is done because it boosts up the weakness of the decision tree.

References

[1] Horie, Takeshi, Masaharu Yajima, and Hiroshi Nakagawa. "Yield forecasting." Agricultural systems 40.1-3 (1992): 211-236.
[2] Crossa, José, et al. "A shifted multiplicative model fusion method for grouping environments without cultivar rank change." Crop Science 35.1 (1995): 54-62.
[3] Liu, Jing, C. E. Goering, and Lei Tian. "A neural network for setting target corn yields." Transactions of the ASAE 44.3 (2001): 705.
[4] Adisa, Omolola M., et al. "Application of artificial neural network for predicting maize production in South Africa." Sustainability 11.4 (2019): 1145.
[5] Awad, Mohamad M. "Toward precision in crop yield estimation using remote sensing and optimization techniques.Portfolio optimization for seed selection in diverse weather scenarios." PloS one 12.9 (2017)" Agriculture 9.3 (2019): 54.
[6] Ma, Yuchi, et al. "Corn yield prediction and uncertainty analysis based on remotely sensed variables using a Bayesian neural network approach." Remote Sensing of Environment 259 (2021): 112408.
[7] Marko, Oskar, et al. ": e0184198.
[8] Heslot, Nicolas, et al. "Integrating environmental covariates and crop modeling into the genomic selection framework to predict genotype by environment interactions." Theoretical and applied genetics 127 (2014): 463-480.
[9] Kross, Angela, et al. "Using artificial neural networks and remotely sensed data to evaluate the relative importance of variables for prediction of within-field corn and soybean yields." Remote Sensing 12.14 (2020): 2230.
[10] LeCun, Yann, Yoshua Bengio, and Geoffrey Hinton. "Deep learning." nature 521.7553 (2015): 436-444.
[11] Ribeiro, Victor Henrique Alves, and Gilberto Reynoso-Meza. "Ensemble learning by means of a multi-objective optimization design approach for dealing with imbalanced data sets." Expert Systems with Applications 147 (2020): 113232.
[12] Tran, Cao Truong, et al. "Multiple imputation and ensemble learning for classification with incomplete data." Intelligent and Evolutionary Systems: The 20th Asia Pacific Symposium, IES 2016, Canberra, Australia, November 2016, Proceedings. Springer International Publishing, 2017.
[13] Ma, Sai, Xianfeng Zhao, and Yaqi Liu. "Adaptive spatial steganography based on adversarial examples." Multimedia Tools and Applications 78 (2019): 32503-32522.