Second International Conference on Nexgen Technologies
Sengunthar Engineering College, Tiruchengode, Namakkal Dist. Tamilnadu (India)
8th - 9th March 2019 www.conferenceworld.in ISBN : 978-93-87793-75-0

Comparative Study of Machine Learning Algorithms for Crop Yield Prediction

P Dhivya[1], Harismitaa R K[2,] Anto Praveena K J[3] , Ashley James[4]

[1] *Assistant Professor, Department of Computer Science and Engineering*
[2,3,4] *Student, Department of Computer Science and Engineering*
SNS College of Technology, Coimbatore (India)

Abstract

Machine learning is an application that focuses on the development of computer programs that can access data and use it learn for themselves. The prime focus of machine learning research is on the development of fast and efficient learning algorithms which can make predictions on data. When dealing with data analytics, machine learning is an approach used to create models for prediction. Developing better techniques to predict crop productivity can assist farmer and other stakeholders in better decision making in terms of agronomy and crop choice. This paper proposes a comparison between Machine Learning algorithms like K-means ,Random Forest, Linear Regression, etc.,

Keywords: Machine Learning ,Data Analytics, cluster analysis, Crop yield, Data mining, K-Means, K-Nearest Neighbor(KNN), Artificial Neural Networks(ANN)

I. INTRODUCTION

Machine learning is an interdisciplinary research area which combines ideas from several branches of science namely, artificial intelligence, statistics, information theory, mathematics, etc. The prime focus of machine learning research is on the development of fast and efficient learning algorithms which can make predictions on data. When dealing with data analytics, machine learning is an approach used to create models for prediction.

Outline of machine learning:

- Machine Learning Applications and Practice :

To give the audience a grasp of the typical uses of and challenges faced by machine learning in practice. Drawing from the industrial expertise of the presenters in large and small machine learning deployments, we develop a prototypical workflow of machine learning projects.

- Systems support for (large scale) Machine Learning : It start with relational algebra inspired systems, MapReduce and its use in Mahout and Spark. It also approaches that are grounded in numerical computation using

VW and graph processing inspired approaches like Pregel and its Open Source implementation Graph and GraphLab.

- Open Research Issues: Approaches such as deep belief networks and graphical models have recently received lots of attention in the machine learning community and even the mainstream media.
- Machine Learning algorithms

- Linear Regression
- Logistic Regression
- Decision Tree
- SVM
- Naive Bayes
- KNN
- K-Means
- Random Forest

In section II, the related work in which yield prediction and disease detection are discussed. In section III, the proposed work in which compare different data mining algorithms with the same parameters. In section IV discussed the conclusion and future work.

II RELATED WORK

The planning process of Crop Yield in Agricultural management needs simple and accurate estimation algorithms.

Existing System

1) Yield Prediction

Yield prediction, one of the most significant topics in precision agriculture, is of high importance for yield mapping, yield estimation, matching of crop supply with demand, and crop management to increase productivity. The aim of the study was to provide growers with yield-specific information to assist them to optimize their grove in terms of profit and increased yield.

2) Disease detection

One of the most significant concerns in agriculture is pest and disease control in open-air and greenhouse conditions. The most widely used practice in pest and disease control is to uniformly spray pesticides over the cropping area. This practice, although effective, has a high financial and significant environmental cost. The study aimed at the accurate detection of these categories for a more effective usage of fungicides and fertilizers according to the plant's needs.

3) Weed detection

Weed detection and management is another significant problem in agriculture. The accurate detection of weeds is of high importance to sustainable agriculture, because weeds are difficult to detect and discriminate from crops. Again, ML algorithms in conjunction with sensors can lead to accurate detection and discrimination of weeds with low cost and with no environmental issues and side effects.

4) Crop Quality

The aim of the study was quality improvement while the minimizing fiber damage. Another study regards pears production and, more specifically, a method was presented for the identification and differentiation. The accurate detection and classification of crop quality characteristics can increase product price and reduce waste.

5) Species Recognition

The last sub-category of crop category is the species recognition. The main goal is the automatic identification and classification of plant species in order to avoid the use of human experts, as well as to reduce the classification time.

III PROPOSED WORK

To compare different data mining algorithms with the same parameters on the 10fold cross validation test to predict the crop yield. The accuracy that breaks all the disadvantages of other algorithms is represented in the fig1.1

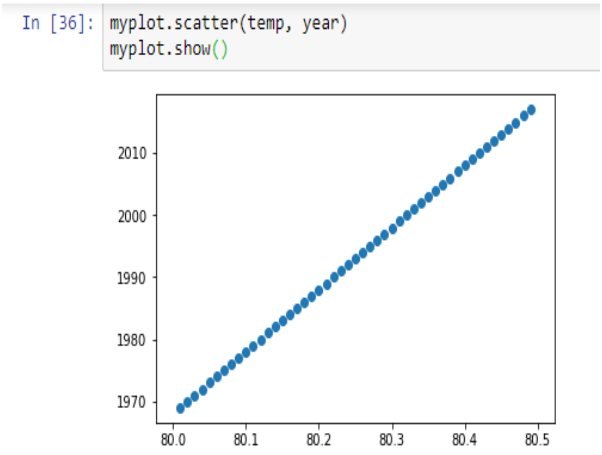

Fig. 1.1 Proposed work in crop yield prediction

Different data mining classification algorithms like K-nearest Neighbor, K-means, Neural Network, Support Vector Machine, Case-based Reasoning, Decision Tree algorithm, etc. are applied for various application of agriculture domain. A comparative study is done by using linear regression, multiple regression, etc., and to find the accuracy of the algorithms. This study first time demonstrates the application of different data mining classification techniques (as discussed above) in the domain of agriculture for yield prediction. The main objective of using information unseen within the database provides the inspiration to the researcher in the area of agriculture for applying such techniques to do forecast for imminent trends of agricultural progressions. For the same, so many works is being done by employing various data mining techniques on agriculture database.

Advantages

- Increased accuracy of the algorithm, also the speed of execution.
- Using algorithms such as fuzzy classifier will automate the process.
- Provides feasible and robust solution for detecting the infected areas
- Forms highly non linear as well as adaptive boundaries.
- Usability to determine the relative influence of predictor variable.
- Flexibility for separating the data

Linear Regression

Linear regression is used to estimate real world values like cost of houses, number of calls, total sales etc. based on continuous variable(s). Here, we establish relationship between dependent and independent variables by fitting a best line.The outcome of linear regression is represented in the fig1.2

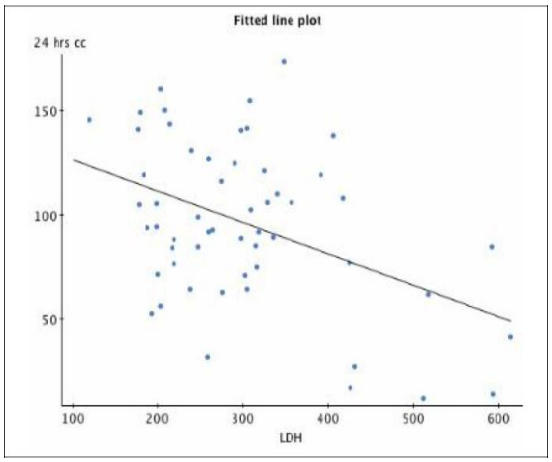

Fig 1.2 Outcome of LinearRegression

Logistic Regression

Logistic regression is another technique borrowed by machine learning from statistics. It is the preferred method for binary classification problems, that is, problems with two class values. It is a classification algorithm and not a regression algorithm as the name says. It is used to estimate discrete values or values like 0/1, Y/N, T/F based on the given set of independent variable(s). It predicts the probability of occurrence of an event by fitting data to a logistic function. Hence, it is also called **logistic regression**. Since, it predicts the probability, its output values lie between 0 and 1.The outcome of logistic regression is represented in the fig1.3

ISBN : 978-93-87793-75-0

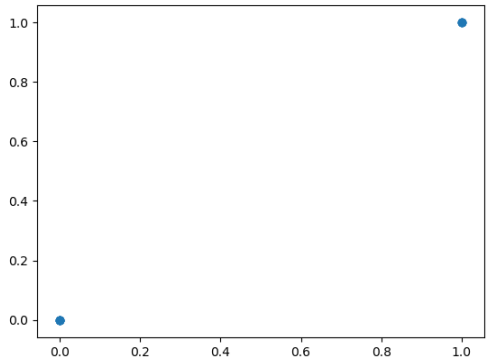

Fig 1.3 Outcome of Logistic Regression

K-means

It is a type of unsupervised algorithm which deals with the clustering problems. Its procedure follows a simple and easy way to classify a given data set through a certain number of clusters (assume k clusters). Data points inside a cluster are homogeneous and are heterogeneous to peer groups. The outcome of K-means is represented in the fig 1.4

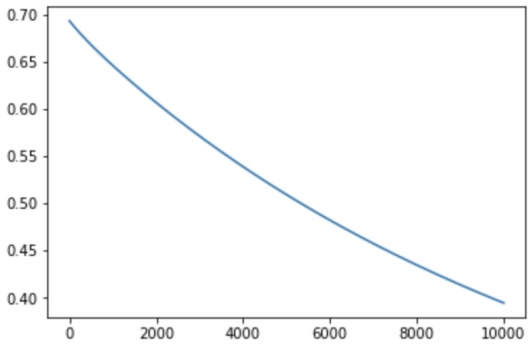

Fig1.4 Outcome of K-means

Random Forest

Random Forest is a popular supervised ensemble learning algorithm. 'Ensemble' means that it takes a bunch of 'weak learners' and has them work together to form one strong predictor. In this case, the weak learners are all randomly implemented decision trees that are brought together to form the strong predictor — a random forest.

The **sklearn.ensemble** module includes two averaging algorithms based on randomized decision trees - **the Random Forest algorithm** and the **Extra-Trees method**. Both algorithms are perturb-and-combine techniques [B1998] specifically designed for trees. This means a diverse set of classifiers is created by introducing randomness in the classifier construction. The prediction of the ensemble is given as the averaged prediction of the individual classifiers. The outcome of Random Forest is represented in the fig1.5

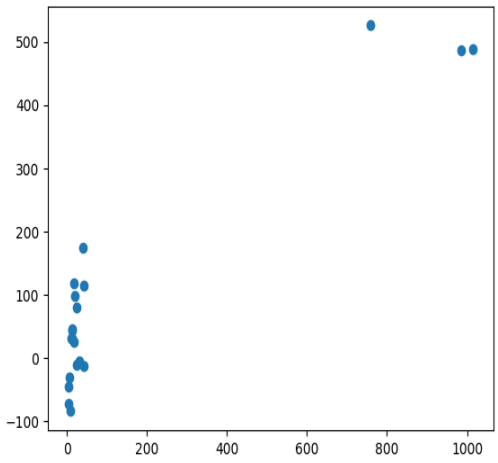

Fig 1.5 Outcome of Random Forest

IV CONCLUSION

We have discussed the prediction of crop yield by comparing some techniques and machine learning algorithms. The main aim of the system is to reduce labor requirements in manual harvesting and handling operations. The study was to provide growers with yield-specific information to assist them to optimize their grove in terms of profit and increased yield. The accuracy of the various algorithms differed from one another and each proved better in different occasion and datasets. The need for an algorithm that will prove efficient every time is the reason that gave rise to the comparative study of machine learning algorithms in crop yield prediction. The future work lies in building a new algorithm that would overcome all the disadvantages of the listed algorithms and to be able to avoid pre-analysis of data that was a hassle in implementing in the current project.

REFERENCES

[1] Samuel, A.L. Some Studies in Machine Learning Using the Game of Checkers. IBM J. Res. Dev. 1959, 44, 206–226. [CrossRef].

[2] Kong, L.; Zhang, Y.; Ye, Z.Q.; Liu, X.Q.; Zhao, S.Q.; Wei, L.; Gao, G. CPC: Assess the protein-coding potential of transcripts using sequence features and support vector machine. Nucleic Acids Res. 2007, 35, 345–349. [CrossRef] [PubMed].

[3] Cramer, S.; Kampouridis, M.; Freitas, A.A.; Alexandridis, A.K. An extensive evaluation of seven machine learning methods for rainfall prediction in weather derivatives. Expert Syst. Appl. 2017, 85, 169–181. [CrossRef].

[4] Rhee, J.; Im, J. Meteorological drought forecasting for ungauged areas based on machine learning: Using long-range climate forecast and remote sensing data. Agric. For. Meteorol. 2017, 237–238, 105–122. [CrossRef].

[5] Aybar-Ruiz, A.; Jiménez-Fernández, S.; Cornejo-Bueno, L.; Casanova-Mateo, C.; Sanz-Justo, J.; Salvador-González, P.; Salcedo-Sanz, S. A novel Grouping Genetic Algorithm-Extreme Learning Machine approach for global solar radiation prediction from numerical weather models inputs. Sol. Energy 2016, 132, 129–142. [CrossRef].

[6] Barboza, F.; Kimura, H.; Altman, E. Machine learning models and bankruptcy prediction. Expert Syst. Appl. 2017, 83, 405–417. [CrossRef].

[7] Zhao, Y.; Li, J.; Yu, L. A deep learning ensemble approach for crude oil price forecasting. Energy Econ. 2017, 66, 9–16. [CrossRef].

[8] Bohanec, M.; Kljaji´c Borštnar, M.; Robnik-Šikonja, M. Explaining machine learning models in sales predictions. Expert Syst. Appl. 2017, 71, 416–428. [CrossRef].

[9] Takahashi, K.; Kim, K.; Ogata, T.; Sugano, S. Tool-body assimilation model considering grasping motion through deep learning. Rob. Auton. Syst. 2017, 91, 115–127. [CrossRef].

[10] Gastaldo, P.; Pinna, L.; Seminara, L.; Valle, M.; Zunino, R. A tensor-based approach to touch modality classification by using machine learning. Rob. Auton. Syst. 2015, 63, 268–278. [CrossRef]

[11] López-Cortés, X.A.; Nachtigall, F.M.; Olate, V.R.; Araya, M.; Oyanedel, S.; Diaz, V.; Jakob, E.; Ríos-Momberg, M.; Santos, L.S. Fast detection of pathogens in salmon farming industry. Aquaculture 2017, 470, 17–24. [CrossRef]

[12] J. Fragni, R.; Trifirò, A.; Nucci, A.; Seno, A.; Allodi, A.; Di Rocco, M. Italian tomato-based products authentication by multi-element approach: A mineral elements database to distinguish the domestic provenance. Food Control 2018, 93, 211–218. [CrossRef]

[13] Pearson, K. On lines and planes of closest fit to systems of points in space. Lond. Edinb. Dublin Philos. Mag. J. Sci. 1901, 2, 559–572. [CrossRef]

[14] Fang, K.; Shen, C.; Kifer, D.; Yang, X. Prolongation of SMAP to Spatiotemporally Seamless Coverage of Continental U.S. Using a Deep Learning Neural Network. Geophys. Res. Lett. 2017, 44, 11030–11039. [CrossRef]

[15] Fisher, R.A. The use of multiple measures in taxonomic problems. Ann. Eugen. 1936, 7, 179–188. [CrossRef]