PREDICTION OF CROP YIELD IN PRECISION AGRICULTURE USING MACHINE LEARNING

S. Vijayasree
Department of Computer Science, Bharathidasan University, India

Abstract

In situations where vast amounts of data are being gathered and published, machine learning is a potential solution. In order to estimate large-scale agricultural yields, we applied machine learning techniques to agronomic principles. A workflow that emphasises consistency, modularity, and reusability serve as a baseline for the project. To ensure accuracy, we worked to create predictors or traits that could be explained and then used machine learning without leaking any information. MCYFS data from the weather, remote sensing, and soil sensors was used to generate new functionalities. Smaller configuration adjustments allow us to handle many different crops and countries with our modular and reusable work flow design. Standard input data and the methodology can be utilised for repeatable tests with repeatable outcomes. It is from these findings that we may go on to refine our algorithms even further.

Keywords:

Machine Learning, Crop Modeling, Crop Yield Forecasting, Crop Simulation

1. INTRODUCTION

Precision agriculture aims to increase agricultural output and quality while minimising operating expenses and environmental impact (PA). Many production characteristics, including as weather, soil conditions, terrain, irrigation, and fertiliser management, have an impact on the potential growth and output. [1–3]. These inputs are critical for huge agricultural fields, which necessitate timely and precise detection via remote and proximate sensing systems in Pennsylvania. Ground-based vehicles, aircraft, satellites, and handheld radiometers can all be used to collect spectral, spatial, and temporal data on the things they observe.

Weed maps are generated using remote sensing techniques such as satellite and aerial multi-spectral scanning, photography and video. For improved water management in irrigated crops, thermal remote sensing by aircraft thermal images has the potential to determine geographic variations of crop moisture status. A wide range of crops, including wheat, corn, and grapevines are now being monitored using this method. Ground-based platforms have been developed for many various PA activities, including mapping soil property data, predicting evapotranspiration and drought stress, and even mapping the locations of undesirable weeds as well as determining crop water and nitrogen status [4].

Some vegetation attributes can be estimated by remote sensing at visible and near-infrared (vis-NIR) wavelengths [5]. The amount of photosynthetic/photoprotective pigments, such as chlorophyll, as well as the leaf area index are included in this measurement. Applicability, representativeness, environment and precision of implementation of more than 100 vegetation indices. Using existing vegetation indices for real-world applications needs careful assessment of their strengths and weaknesses as well as the unique area in which they will be used. Remotely sensed vegetation indices have been used to estimate agricultural yields. There has been progress in the use of wireless sensor networks and algorithms for integrating data from these networks in PA [6].

Growth and development of plants rely heavily on nitrogen (N), which is closely linked to the photosynthetic process. N, on the other hand, has a significant impact on both the environment and the economy. As a result, spectrometric investigations have been conducted on the optimization of fertilisation for various crops. Destructive and non-destructive methods of determining the plant N status are available. Kjeldahl method of destructive measurement is most commonly used, but chemical analysis is more time-consuming, expensive and labor-intensive [8].

Plant N status can be monitored non-destructively via optical remote sensing, which measures canopy reflectance at visible and near-infrared wavelengths (400–900 nm). To reduce the time and cost of field sample collection, preparation, and laboratory analysis, this measurement is done in-situ, which reduces the number of field samples required. Hyperspectral data from remote sensing has been used to create spectral indices indicative of plant N status [9]-[11] from hyperspectral data sets.

The use of remote sensing in a variety of fields, including geology, forestry, and others, has resulted in massive data accumulation. The amount of data is always increasing, making it impossible for any one person to properly integrate, analyse, and make decisions based on all of it. When data is not homogeneous, such as when it is collected by sensors with varying spatial, temporal, and spectral modalities [12], this is especially true. When it comes to discovering patterns and rules in massive datasets, machine learning (ML) is an emerging technique.

Because of the direct linkage in fertiliser management decisions, crop yield forecast and N status estimation are considered jointly here. It is common practise to use crop yield goals when determining N requirements, both before and during the growing season. An estimation of both would be useful for developing possible site-specific management plans for N fertiliser, especially during the growing season. Here, we'll explain how various machine learning approaches can be used to solve a variety of connected problems. A review of recent works incorporating various ML approaches into agricultural production prediction and N status estimate is presented. Comparative analyses of ML approaches applied to the identical job in PA are also included in this book. We'll go through some of the specifics of the machine learning methods employed in the research we'll be looking at.

2. BACKGROUND

Crop yield optimization at the lowest possible cost while maintaining a healthy ecosystem is one of the most important

objectives in agricultural production. There are several crop management and economic decisions that depend on yield estimation, therefore it critical to catch problems related with crop yield constraints early and address them effectively.

Combinations of red, green and infrared wavelengths are used to calculate vegetation indexes (VIs). They are made to discover functional connections between the traits of crops and the data collected through remote sensing. It been possible to create numerous new vegetation indexes since the introduction of the SR and NDVI, like the enhanced vegetation two-band index (EVI2) and the normalised difference water index (NDWI), to name just a few [15]-[17]. Since so many indicators are available, it is necessary to choose and combine them effectively in order to get the most accurate estimate of crop production possible.

3. PROPOSED MODEL

Because of their complexity, ML algorithms are extremely expensive to develop, repair, and maintain. To better anticipate crop yield (mustard, wheat), machine learning algorithms merged input and output data. It was not possible for the regression model to accurately anticipate extreme values or nonlinear data because of the linear character of the parameters. As a result of the nonlinear and highly adaptive challenges in KNN, existing K-NN models were employed for classification for yield prediction. An increase in the input vector dimensions made it difficult to correctly classify them. Because of the limited amount of data available to estimate crop output, a suitable judgement could not be made during classification.

Different feature groups associated with soil information were examined through the study of the studies such soil maps and types and production areas. Soil maps will show the types of nutrients in the soil as well as the locations where the soil may be located. The crop information characteristics include crop density, growth process in terms of weight, and leaf area index for crops such as mustard crops, wheat, rice, and tomato plants. Humidity, rainfall, precipitation, and forecaster rainfall are all examples of weather features. The nutritional components play a significant influence in the context of various environmental conditions. Nitrogen, potassium, magnesium, zinc, boron, etc., are among the nutrients. Temperature and radiation (gamma), shortwave radiation, solar radiation, and degree days are used to calculate features in the solar information. Wind speed, images, and pressure are all computed using fewer features.

Specifically, we used supervised regression (see Section B of Supplement 1) to estimate crop yields. Learning a function that connects features to labels is the goal of supervised learning, which uses training examples that include both features and labels, like yield statistics. In order to train and test, we divided the dataset into two halves. We expanded the test set by including the most recent few years for each location when using the yield trend. It was necessary to impose this restriction since yield trend estimates from previous years would be included in following years, and this would lead to data leakage. We may have utilised random splits instead of the yield trend. A comparison with MCYFS required the same test years in all regions.

So every nth year we added to the test set, with n determined by how many times we tested We allotted 70% of the data for training, and 30% for testing, in both situations. The training set was used to develop and test a model, while the test set served as the final assessment. Our crop calendar and indicator statistics were derived solely from the training data.

By separating the training set into validation folds, we were able to tune the hyperparameters of feature selection (the number of features to pick) and prediction algorithms (e.g., the number of neighbours for k-nearest neighbours). We were unable to conduct cross validation using the yield trend because the test fold could end up in a bin before the training folds and that would result in information leakage. As a result, we employed a k-fold sliding validation that was time-based.

As an example, data from 1994 to 2018 was available for NL, and the training years were 1994 through 2011. K-fold sliding validation was trained using data from 1994 to 2007 for the first iteration, followed by data from 1995 to 2008 for iteration 2, etc. until the fifth iteration, which was tested on data from 1998 to 2011. We used k-fold cross-validation when yield trend was not used.

In order to prevent data leaking during the feature selection and training stages, we developed pipelines consisting of scaling, selection, and training stages. The pipelines guaranteed that just the training data was used for each stage of training and optimization. The parameters for scaling features, the number of features to pick, and the feature weights for the trained model were learned from the training set in this manner. " In addition, we tuned the hyperparameters using only the training set. The pipeline was executed for each iteration of 5-fold sliding validation or 5-fold cross-validation when optimising hyperparameters. Because of this, all phases of the pipeline (feature scaling, feature selection, and training) were conducted with the training folds, and the trained model was assessed with the matching test folds.

In terms of modularity, we aimed to make it as simple as possible to improve and extend the baseline. We reduced the amount of interdependencies among the various stages of the process. In order to keep the feature design process as flexible as possible, we used extensible data structures. Designing new features or upgrading current features with fresh data was the purpose of the project. For example, extreme weather features count the number of days that fall outside of the average. The method is generic and repeatable because of the usage of averages and standard deviations of indicators. Crop-specific thresholds for distinct indicators can be used to manually design more accurate and predictive features if they are available.

It was important to us to build the workflow so that it could be used for various crops and nations. Standardizing filenames, file formats, and data columns through data homogenization allowed us to reduce the quantity of input needed to conduct the workflow. We applied the same design concepts to a variety of diverse projects. To run the workflow for multiple crops, countries, and NUTS levels, data homogenization and configuration choices were necessary. To avoid having to provide all of the options for each experiment, we set most of the setup options to appropriate defaults.

3.1 PERFORMANCE EVALUATION

The modulation factor values of ML algorithms vary depending on the different crop feature divisions. ANN is used

when the number of input elements is minimised. In order to obtain an accurate estimation of crop yield, the best feature was empirically chosen. Linear functions in large output sample spaces can be cumbersome, and sophisticated optimizations can be reduced to basic linear function optimization utilising ML technique regression. An ML algorithm can be used to estimate crop yields by using a large dataset of soil samples. Farmers were able to significantly boost crop yield with the use of machine learning techniques that were applied to field observations in the agriculture sector.

Table.1. Comparison of Classification Accuracy

Performance Metrics	**CNN**	**LSTM**	**DNN**	**KNN**
Accuracy	78.33	74.47	79.06	78.48
Precision	86.07	81.84	86.89	86.24
Recall	71.86	68.32	72.54	72.00
F-Measure	66.89	63.60	67.52	67.02
Computation time	0.86	0.82	0.87	0.87

The diversity of characteristics covered in this study is mostly based on the availability of data, and each of the studies will investigate PA using ML techniques that differ from the features. Because the data-set availability was a major factor in determining which features to include in the model, the more features included were not always more effective. As a result, only the most effective features were chosen for further testing.

A wide range of machine learning (ML) approaches were applied to provide the best predictions for PA, including neural networks, random forests, KNN regression, and more. CNN, LSTM, and DNN are the most commonly utilised algorithms in PA, however there is still room for improvement. In this study, a number of existing models for predicting crop yields, such as temperature and meteorological conditions, are examined.

After everything was said and done, the results of the experiment demonstrated that applying ML to the agricultural domain helped advance crop prediction. However, there was still room for improvement in the selection of features affecting agricultural production as a result of temperature variance. Studying the most important possibility, such as firstly the delay to border topographical areas, necessitated further explicit treatment.

It then uses machine learning algorithms and features from deterministic crop models to determine the best statistical CO2 fertilisation for the model... The crop yield estimation can be enhanced by future research if the above-mentioned aims are followed. Fertilizer should also be taken into account when calculating crop yields in order for agriculturalists to make a better decision in the event of poor crop yield estimation. We need to create and develop a model for PA based on the results of this study.

4. CONCLUSION AND FUTURE WORK

Sensing and machine learning (ML) have come a long way in the recent decade. It is expected that these improvements will continue to provide cost-effective and more complete datasets, coupled with more complex algorithmic solutions, that will allow for better crop and environment status estimates and decision-making. There is a promising path ahead of us that has the potential to revolutionise crop output management. It has already been proven that ML approaches can be applied to a wide range of PA.

Algorithms and sensors will continue to evolve in tandem, resulting in the following future trends: Optimized, focused use of sensors and existing ML algorithms for specific PA activities The incorporation of expert information into ML approaches targeted at modelling and decision making in various elements of PA is interconnected. The integration of several ML and signal processing approaches into hybrid systems to profit from the strengths of those techniques and compensate for their respective weaknesses. Fusion of information from sensors with diverse spatial and spectral resolutions and properties. It is possible to acquire active optimal data, combine information, and update models for high value locations using a dynamic combination of stationary and mobile equipment.

REFERENCES

[1] R. Wanjari, K.G. Mandal, P. Ghosh, T. Adhikari and N.H. Rao, “Rice in India: Present Status and Strategies to Boost Its Production Through Hybrids”, *Journal of Sustainable Agriculture*, Vol. 28, No. 2, pp. 19-39, 2006.

[2] K. Thiyagarajan and R. Kalaiyarasi, “*Status Paper on Rice in Tamilnadu*”, Rice Knowledge Management Portal Publisher, 2010.

[3] R. Rupnik, M. Kukar, P. Vracar and Z. Bosnic, “AgroDSS: A Decision Support System for Agriculture and Farming”, *Computers and Electronics in Agriculture*, Vol. 161, pp. 260-271, 2019.

[4] G. Yogeswari and A. Padmapriya, “Recommender System for Nutrient Management Based on Precision Agriculture”, *International Journal of Recent Technology and Engineering*, Vol. 8, No. 4, pp. 1-12, 2019.

[5] Venkatalakshmi Balakrishnan, “Decision Support System for Precision Agriculture”, International Journal of Research in Engineering and Technology, Vol. 3, No. 19, pp. 849-852, 2014.

[6] Guidelines for Rice, Available at: http://www.knowledgebank.irri.org/decision-tools/ricedoctor, Accessed on 2020.

[7] Production and Irrigation Statistics, Available at: https://visualize.data.gov.in/all_visualization, Accessed on 2020.

[8] Water Related Data, Available at: https://www.indiawaterportal.org/datafinder, Accessed on 2020.

[9] R. Ruba Mangala and A. Padmapriya, “Visualizing the Impact of Climatic Changes on Pest and Disease Infestation in Rice”, *International Journal of Recent Technology and Engineering*, Vol. 8, No. 3, pp. 8413-8421, 2019.

[10] M. Abedinpour, A. Sarangi, T.B.S. Rajput, M. Singh and T. Ahmad, “Performance Evaluation of Aqua Crop Model for Maize Crop in a Semi-Arid Environment”, *Agricultural Water Management*, Vol. 110, pp. 55-66, 2012.

[11] M. Sujaritha, S. Annadurai, J. Satheeshkumar, S. Kowshik Sharan and L. Mahesh, “Weed Detecting Robot in Sugarcane Fields using Fuzzy Real Time Classifier”,

Computers and Electronics in Agriculture, Vol. 134, pp.160-171, 2017.

[12] Michelle Araujo E Viegas, Avinash Kurian, Victor Joshua Rebello and Niraj Mangaldas Gaunker, "Weed Detection using Image Processing", *International Journal for Scientific Research and Development*, Vol. 4, No. 11, pp. 660-662, 2017.

[13] Na Zhao, Shi Kai Sui and Ping Kuang, "Research on Image Segmentation Method Based on Weighted Threshold Algorithm", *Proceedings of International Computer Conference on Wavelet Active Media Technology and Information Processing*, pp. 307-311, 2015.

[14] Nishant Deepak Keni and Rizwan Ahmed, "Neural Networks based Leaf Identification using Shape and Structural Decomposition", *Proceedings of International Conference on Global Trends in Signal Processing, Information Computing and Communication*, pp. 225-229, 2015.

[15] R. Aravind, M. Daman and B.S. Kariyappa, "Design and Development of Automatic Weed Detection and Smart Herbicide Sprayer Robot", *Proceedings of IEEE Conference on Recent Advances in Intelligent Computational Systems*, pp. 257-261, 2015.

[16] J. Behmann, A.K. Mahlein, T. Rumpf, C. Romer and L. Plumer, "A Review of Advanced Machine Learning Methods for the Detection of Biotic Stress in Precision Crop Protection", *Journal of Precision in Agriculture*, Vol. 16, pp. 239-260, 2015.

[17] G. Bhanumathi and B. Subhakar, "Smart Herbicide Sprayer robot for Agriculture Fields", *International Journal of Computer Science and Mobile Computing*, Vol. 4, No. 7, pp. 571-574, 2015.