


RESEARCH ARTICLE | NOVEMBER 21 2023

Application of data mining and machine learning in food and agriculture industry towards precision agriculture

Thanwamas Kassanuk; Khongdet Phasinam 



AIP Conf. Proc. 2587, 050013 (2023)

<https://doi.org/10.1063/5.0150472>



CrossMark

AIP Advances

Why Publish With Us?

-  **25 DAYS**
average time to 1st decision
-  **740+ DOWNLOADS**
average per article
-  **INCLUSIVE**
scope

[Learn More](#)



Application of Data Mining and Machine Learning in Food and Agriculture Industry towards Precision Agriculture

Thanwamas Kassanuk and Khongdet Phasinam^{a)}

Faculty of Food and Agricultural Technology, Pibulsongkram Rajabhat University, Thailand.

^{a)}Corresponding author: phasinam@psru.ac.th

Abstract. It is imperative that agricultural productivity keep pace with population growth due to the limited supply of natural resources. The primary goal here is to increase productivity even in the face of adverse environmental conditions. As a result of advances in agricultural technology, precision farming is increasingly being used to increase yields. Machine learning encompasses a wide range of techniques that may be used to learn predictive rules from previous data and develop a model that can anticipate unknown future information. A predictive model may be built using machine learning by analyzing data samples to detect trends and creating decision rules. Smart agriculture is a modern agricultural paradigm that evaluates the farm as a collection of tiny units and identifies irregularities in output and demand for individual units.. It is the ultimate objective of smart agriculture to minimize agricultural costs in order to boost profits. Farming strategies that are cutting-edge are used by smart farmers. A machine learning algorithm's ability to forecast yields makes farming more efficient and effective.

INTRODUCTION

It is imperative that agricultural productivity keep pace with population growth due to the limited supply of natural resources. The primary goal here is to increase productivity even in the face of adverse environmental conditions. As a result of advances in agricultural technology, precision farming is increasingly being used to increase yields. Methods for learning predictive rules from prior data and building a model that can predict unknown future data fall under the umbrella of machine learning (ML). When a computer can learn from data, it is called machine learning. When machine learning and Bigdata methods are employed, robots may be combined with artificial intelligence and act and think like humans without being explicitly taught. Thus, we may say that Bigdata and data mining are mutually reinforcing and mutually reinforcing. [2]

However, most datasets obtained from multiple sources are erroneous, which is a problem for machine learning algorithms. Machine learning algorithms have a number of difficulties when dealing with data that is missing, duplicated, outliers, or otherwise inconsistent.

According to computer science, self-training systems are known as machine learning, and they may learn from a given data set without being explicitly programmed. It's a collection of strategies that enables a computer program to accurately forecast the outcome. It is the purpose of machine learning when fresh data enters a computer system to construct an algorithm that predicts an output using statistical formulas. As a result, machine learning analyzes data samples in order to uncover patterns and develop decision rules for constructing a predictive model that can be used to forecast future data. The more experience they gain, these predictive models may be able to learn on their own and make decisions based on specific circumstances without the need for human interaction. Both supervised and unsupervised machine learning exist. There are a lot of algorithms in each field. [3]

MACHINE LEARNING TECHNIQUES AND PREDICTIVE ANALYSIS

A Study of Machine Learning Techniques

Algorithm Iterative Dichotomizer-3 (ID3) The ID-3 method (Iterative Dichotomizer-3) was developed by J. Ross Quinlan and is the first evolving decision tree based system. The entropy and information gain measurements are used in this method. In the original dataset, the entropy measure of the functional characteristics is computed for each of the iteration. Datasets are separated into subsets depending on the property with the lowest entropy and highest information gain (split attribute). Recursively repeating the method on each subset of data until it is precisely categorized to its target classes. Decision trees are created using non terminal nodes, and the ultimate subset of a branch specifies the terminal nodes. The non terminal node is defined by the split property, while the terminal node is the class labels. ID-3-based decision tree model is used to categorize and forecast cardiac disease at an earlier stage. Using well-known decision tree techniques like CART and ID-3, a prediction model with a large health dataset is built. Validation is carried out using a 10-fold cross validation technique. The results show that decision tree classification approaches may be used to build an accurate and efficient model of prediction models. Using the ID-3 technique, smaller datasets may be processed more efficiently, resulting in superior results. The problem arises when dealing with large amounts of data, such as electronic health records, which are constantly updated. Using ID-3 based decision tree classification algorithms, the accuracy of health data categorization is increasingly affected by model over fitting. [4]

The decision tree classification model is described by the author in [5]. The first step is to build a classification model that accurately categorizes the data. Classification is done from the root node to the leaf nodes using a top-down technique. The entropy measurements are used to pick an efficient attribute and to divide the provided data into subsets. Decisions are based on the attribute with the greatest normalized information gain. It is efficient because it removes unnecessary properties from the branches. It's also good at handling characteristics with numeric values and missing data. With numeric characteristics, the decision tree gets more difficult.

Decision tree C-5 was developed to address some of the shortcomings of the C4.5 method, according to researchers in [6]. It allows for quicker computations while consuming less storage space and reducing the size of the tree. In addition to supporting boosting, it also automatically eliminates any traits that aren't relevant to the categorization. Limits are placed on the training dataset to ensure that the search is focused on relevant information. A second series of tests is used to verify the final categorization findings. The application of association rules efficiently ties the risk factor for heart disease to the severity of the condition. The C-5 decision tree approach decreases the number of association rules and improves the accuracy of the results.

Classification and regression trees, as described in [7], are a well-known approach of decision tree-based classifiers, as the author describes in his work. There are distinct inputs and base points for each base node. Numeric values are assumed for the input attribute value. Predictions are made using the output variable represented by the leaf node. A statistical model is built using discriminant analysis to categorize the dataset more accurately. In terms of categorical and continuous qualities, it is a good fit.

Ensemble learning, according to authors [8], is capable of doing classification and regression. During the training phase, it builds a large number of decision trees and uses regression methods to predict the outcomes of each tree. It has a lower standard deviation and can more readily link the various characteristics of the input data in order to make predictions. The random forest classification algorithms are difficult to understand at first, which is why there was a lack of enthusiasm for them.

Bayesian classification is a statistical method of classification based on the bayes theorem, according to authors [9]. It assumes a probabilistic model and works well with both category and numerical data.. Compared to other classification methods, this one is a lot quicker. With the help of a set of standard procedures and guidelines, it creates the categorization model. It presupposes that each and every one of the system's attributes is a part of it. These are commonly used classification algorithms that can handle even the most complicated and big datasets.

The most reliable technique for pattern recognition and data categorization, according to research [10], is the K-nearest neighbor approach. The distance functions or similarity measure is the foundation of K-nearest neighbor algorithms. Classifying new instances based on similarity is done by storing their current state. Instance-based learning is used for efficient categorization. According to the majority of its neighbors, a new dataset instance is categorized. Data from both training and testing sets is used to compute the distance between the two sets to begin the process, the first step is to determine the value of k and determine the distance between the instances.

PREDICTIVE ANALYSIS

A good model for predicting the results of the input datasets is built using predictive analytic techniques. With the help of proper data mining techniques, rules, and relationships, it is possible to build predictive models. It enhances the data's value and predicts how an event will play out in the future. During predictive analysis, classification is the process of properly classifying and mapping the supplied input data. It is a form of data analysis known as predictive analysis that uses data to anticipate future occurrences. Uses statistical and machine learning concepts as well as data mining and artificial intelligence to create predictions. It analyzes raw or unprocessed data and discovers the behavioral patterns connected with it to extract useful information. Individual elements' prediction probabilities are quantified using this metric. It is possible to use this method to make educated guesses about what will happen in the future. It relies on the history of projected values and dependant variables to forecast the unknown outcomes. Using prescriptive analytic approaches, it can increase the system's ability to make better decisions.

Predictive analysis is becoming increasingly dependent on machine learning approaches, according to a study [11]. In order to make predictions, predictive analysis follows a five-step method. The first step is to determine the scope of the work. First, the data sets needed to make predictions are determined, and then the scope of the project and its goals are laid out. In the second stage, data is gathered from a variety of sources. Data is cleaned to eliminate any noise or irrelevant information before any useful information can be gleaned from it. The third step is to use statistical models to test the assumptions and hypotheses. A predictive model is constructed once the assumptions are proven to be correct. The deployment procedure is the last stage. The prediction model is then used in real time to generate findings that may be used to make better-informed choices. Model monitoring is the fifth stage of predictive analysis. The prediction model's performance is constantly evaluated throughout this phase. Predictive analysis may be carried out more quickly and effectively if these seven procedures are followed.

APPLICATION OF MACHINE LEARNING IN FOOD AND AGRICULTURE INDUSTRY

Machine learning techniques are used in object detection, classification, grading and sorting of various fruits and crops.

Object Detection and Classification

Weeds and cultivated crops may be clearly distinguished in field photos using artificial neural networks [9]. Corn plant and weed characteristics were extracted in intensity form from digital photographs, and a neural network trained to categorize the images as weed and corn plant were used in this initial part of the process. In order to distinguish between the weed and maize plants, the neural network has to be properly trained. The trial was a complete success, with a success rate of 100%.

To estimate soybean and maize production in Maryland, an artificial neural network with a back propagation algorithm and a variable learning rate and number of neurons in the hidden layer was created. According to this study, using an ANN model to anticipate soybean and corn yields was more accurate and reliable than using a regression model, which was also found to be more accurate.

Grading and Sorting

An essential part of the quality evaluation process is the grading of the items. It not only distributes grades, but it also determines the monetary and time worth of the delivered work. The process of grading necessitates both specialized knowledge and a close examination of the provided output. There are several factors that go into determining a product's quality. Demand for high-quality items is strong as well as their return rates. As the final but most critical phase in the production process, quality assurance is essential for every manufacturing or agricultural operation.

Using online fruit grading, researchers attempted to classify apple fruit into four grade groups. The study looked at the apple's outward traits to determine its quality, and the categorization was about 90% accurate. Horticulture items are heavily invested in by the fruit market. In this domain, there may be more studies on quality rating. A number of fruit grading methods have previously utilized machine vision, according to the literature.

According to European standards, author [12] did a research on apple fruit grading and chose one of four categories: 'I, II, 'Extra, and 'Rejected.' The characteristics of the apple, such as its color, blemish, and size, were used to classify it. Golden Delicious apples and Jonagold apples were both classified as having a success rate of 78% and 72% using the categorization techniques used in this investigation on the two kinds. To evaluate apples, a fuzzy logic system was used in another study.

Crop Disease Detection

Computer vision, image processing, and machine learning techniques are used in the automated leaf disease diagnostic system to analyze photographs of diseased leaves. The farmer can make an educated choice regarding a plant illness thanks to automated disease detection equipment that speeds up the diagnosis process. Before, the pathologist had to confirm the condition by sending the diseased leaf to a pathology lab, which was a time-consuming process. Crop yields drop as a result of the delayed response. As a result, disease identification must be automated to boost agricultural yields. [13][14] [15]

CONCLUSION

Due to the scarcity of natural resources, agricultural output must maintain pace with population increase. The key objective here is to boost productivity even in the face of challenging environmental circumstances. Precision farming is increasingly being utilized to boost yields as agricultural technology progresses. Machine learning refers to a variety of approaches that may be used to learn predictive rules from prior data and create a model that can predict unknown future information. Machine learning may be used to create a prediction model by evaluating data

samples to find trends and establishing decision rules. The ultimate goal of smart agriculture is to reduce agricultural costs in order to increase profits.. Cutting-edge farming methods have been adopted by the rich and famous. The predictive nature of machine learning algorithms enables smart farming. This article provides an in depth study of various machine learning techniques for classification of available agriculture data towards making smart farming possible. Applications of machine learning in agriculture are also discussed in detail. This study will help future researchers in making predictive model for precision agriculture.

REFERENCES

- [1] K. N. Bhanu, H. J. Jasmine and H. S. Mahadevaswamy, "Machine learning Implementation in IoT based Intelligent System for Agriculture," 2020 International Conference for Emerging Technology (INCET), 2020, pp. 1-5, doi: 10.1109/INCET49848.2020.9153978.
- [2] A. Sharma, A. Jain, P. Gupta and V. Chowdary, "Machine Learning Applications for Precision Agriculture: A Comprehensive Review," in *IEEE Access*, vol. 9, pp. 4843-4873, 2021, doi: 10.1109/ACCESS.2020.3048415.
- [3] A. Muniyasamy, "Machine Learning for Smart Farming: A Focus on Desert Agriculture," 2020 International Conference on Computing and Information Technology (ICCIT-1441), 2020, pp. 1-5, doi: 10.1109/ICCIT-144147971.2020.9213759.
- [4] P. S. Nishant, P. Sai Venkat, B. L. Avinash and B. Jabber, "Crop Yield Prediction based on Indian Agriculture using Machine Learning," 2020 International Conference for Emerging Technology (INCET), 2020, pp. 1-4, doi: 10.1109/INCET49848.2020.9154036.
- [5] M. Santhosh, M. D. Sai and S. Mirza, "Ensemble deep learning model for wind speed prediction," 2020 21st National Power Systems Conference (NPSC), 2020, pp. 1-5, doi: 10.1109/NPSC49263.2020.9331836.
- [6] M. Utsumi, I. Shigemori and T. Watanabe, "Forecasting Electricity Demand with Dynamic Characteristics Based on Signal Analysis and Machine Learning," 2020 59th Annual Conference of the Society of Instrument and Control Engineers of Japan (SICE), 2020, pp. 1049-1054, doi: 10.23919/SICE48898.2020.9240281.
- [7] S. Nandhini and J. Marseline K.S., "Performance Evaluation of Machine Learning Algorithms for Email Spam Detection," 2020 International Conference on Emerging Trends in Information Technology and Engineering (ic-ETITE), 2020, pp. 1-4, doi: 10.1109/ic-ETITE47903.2020.312.
- [8] C. M. Suneera and J. Prakash, "Performance Analysis of Machine Learning and Deep Learning Models for Text Classification," 2020 IEEE 17th India Council International Conference (INDICON), 2020, pp. 1-6, doi: 10.1109/INDICON49873.2020.9342208.
- [9] C. Yang, S. O. Prasher, J. Landry, H. S. Ramaswamy, and A. Ditommaso, "Application of artificial neural networks in image recognition and classification of crop and weeds," *Can. Agric. Eng.*, vol. 42, no. September, pp. 147–152, 2000.
- [10] M. Kaul, R. L. Hill, and C. Walthall, "Artificial neural networks for corn and soybean yield prediction," *Agric. Syst.*, vol. 85, no. 1, pp. 1–18, Jul. 2005.
- [11] V. Leemans, H. Magein, and M.-F. Destain, "On-line Fruit Grading according to their External Quality using Machine Vision," *Biosyst. Eng.*, vol. 83, no. 4, pp. 397–404, Dec. 2002.
- [12] J. Faria, T. Martins, M. Ferreira, and C. Santos, "A computer vision system for color grading wood boards using Fuzzy Logic," *2008 IEEE Int. Symp. Ind. Electron.*, pp. 1082–1087, Jun. 2008.
- [13] S. Minaei, M. Jafari, and N. Safaie, "Design and development of a rose plant disease-detection and site-specific spraying system based on a combination of infrared and visible images," *Journal of Agricultural Science and Technology*, vol. 20, no. 1, pp. 23–36, 2018.
- [14] J. Qin, T. F. Burks, M. A. Ritenour, and W. G. Bonn, "Detection of citrus canker using hyperspectral reflectance imaging with spectral information divergence," *Journal of food engineering*, vol. 93, no. 2, pp. 183–191, 2009.

[15] S. Thomas, M. T. Kuska, D. Bohnenkamp, A. Brugger, E. Alisaac, M. Wahabzada, J. Behmann, and A.-K. Mahlein, "Benefits of hyperspectral imaging for plant disease detection and plant protection: A technical perspective," *Journal of Plant Diseases and Protection*, pp. 1–16, 2018.