Inquiry on Various Data Mining approaches for Crop Yield Prognosis

Shresta S, PG Scholar
Department of Computer Science and Engineering
Vidyavardhaka College of Engineering
(E-mail: shrestas19@gmail.com)

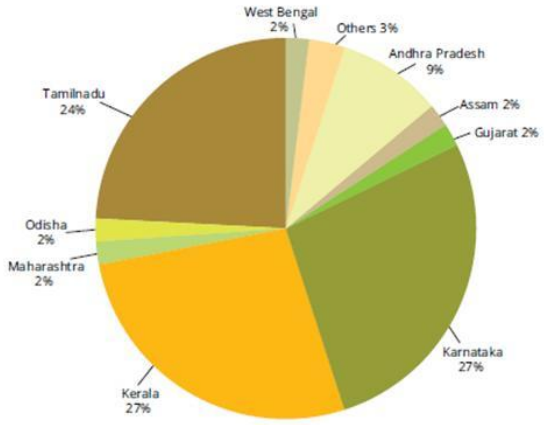
Natesh M, Associate Professor
Department of Computer Science and Engineering
Vidyavardhaka College of Engineering
(E-mail: natesh.m@vvce.ac.in)

***Abstract*—**India is a country where agriculture and agriculture related industries are the major source of living for the people. Agriculture is a major source of economy of the country. It is also one of the countries which suffer from major natural calamities like drought or flood which damages the crop. This leads to huge financial loss for the farmers thus leading to the suicide. Predicting the crop yield well in advance prior to its harvest can help the farmers and Government organizations to make appropriate planning like storing, selling, fixing minimum support price, importing/exporting etc. Predicting a crop well in advance requires a systematic study of huge data coming from various variables like soil quality, pH, EC, N, P, K etc. As Prediction of crop deals with large set of databases thus making this prediction system a perfect candidate for application of data mining. Through data mining we extract the knowledge from the huge size of data. This paper presents the study about the various data mining techniques used for predicting the crop yield. The success of any crop yield prediction system heavily relies on how accurately the features have been extracted and how appropriately classifiers have been employed. This paper summarizes the results obtained by various algorithms which are being used by various authors for crop yield prediction, with their accuracy and recommendation.

***Keywords*—** *crop yield prediction, data mining, data mining algorithm, models, accuracy and Recommendation.*

I INTRODUCTION

India is a land of agriculture-based country where most of the people derive their living from this sector. Agriculture is having a great impact on the country's economy. In the last decade India has seen serious natural calamities like drought or flood. Due to such disasters there is a huge loss to crop production and ultimately to the farmers. Due to such financial loss many farmers are committing suicide. If natural calamities are not present, then there may be sudden pest attack destroying the crop. In any case farmer and the crop are always at the edge of risk. Government policies are there but that is not enough. In order to overcome from this issue many research institutes and students using different emerging application domains and technologies and have introduced various new constraints and methods for information technology. Information technology has become more and more a part of our day to day life, especially true for agriculture and other fields as well. Figure 1. shows the major crop producing states of India.



Source: Ministry of Agriculture

Figure 1: Major crop producing states

Prediction of crop yield in advance can help the farmers and the Government bodies to plan for storage, selling, and fixing minimum support price, importing/exporting etc. Figure 2 shows an example of the Redgram crop with seed buds.

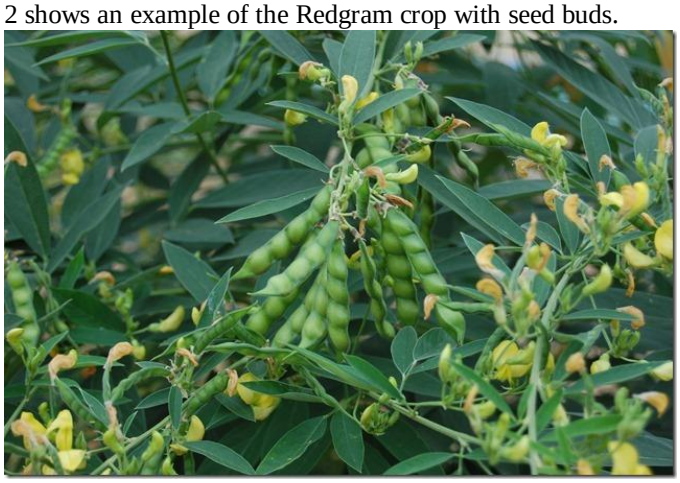


Figure 2 Redgram crop seed buds

Information technology can be used to avert the risk usually associated with the agriculture and it can also be used to predict the crop yield more accurately prior to harvest .Yield prediction needs different kinds of data gathered from different sources like meteorological data, Agri-meteorological, soil (pH, N, P, K) data, remotely sensed data, agricultural statistics etc.[1].To handle such a huge data the best option we have is Data Mining. Data mining is a method by which one can extract the knowledge from the huge bulk of data.

Data Mining techniques are mainly divided in two groups,

- Classification
- Clustering techniques

Classification techniques are designed for classifying unknown samples using information provided by a set of classified samples. This set is usually referred to as a training set as it is used to train the classification technique how to perform its classification. Generally, Neural Networks and Support Vector Machines, these two classifications techniques learn from training set how to classify unknown samples.

Data mining process is separated into seven methods [13]:

- Data cleaning.
- Data integration.
- Data selection.
- Data transformation.
- Data mining.
- Pattern estimation.
- Knowledge display.

II. METHODOLOGY

Through crop yield prediction system better planning and decisions can be chalked out for enhancing the yield

A. Input

Most of the research papers that were studied have considered some climatic parameters like temperature, humidity, rainfall. Some agronomical parameters like soil, nutrient contents like N, P, K, and pesticides etc. The values of these variables have been taken as input.

B. Preprocessing (Noise Removal)

For the successful application of data mining a huge set of datasets is required. The data which is acquired from various resources are sometime in raw form. It may contain some incomplete, redundant, inconsistent data. Therefore, in this step such redundant data should be filtered. Data should be normalized.

C. Feature Extraction (Attribute Selection)

This step aims at identifying and using most relevant attribute from the dataset. Through this process irrelevant and redundant information is removed for the application of classifiers.

D. Output

The output is the crop yield prediction per acre with some of the recommendation. A broad outline of the crop prediction approach is shown in Figure 3.

Another classification technique, K-Nearest Neighbor [10], does not have any learning phase, because it uses the training set every time a classification must be performed. A training set is known, and it is used to classify samples of unknown classification. The basic assumption in the K-Nearest Neighbor algorithm is that similar samples should have similar classification. The parameter K shows the number of similar known samples used for assigning a

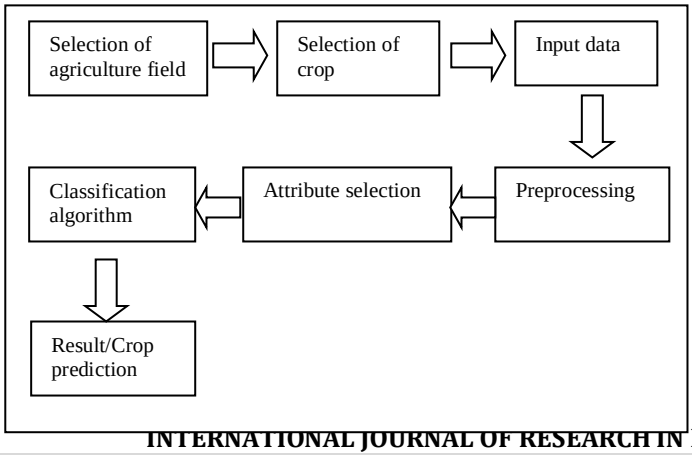


Figure 3 crop prediction system

A Brief overview of the Crop Prediction System:

- Selection of agriculture field: Consider any agriculture field for the crop prediction system.
- Selection of crop: consider any crop of choice which will be sown in that field.
- Input data: Data may include information regarding soil (Nitrogen (N), Phosphorus(P), Potassium(K) content), Micronutrients present in soil, Moisture in soil etc. which is collected over some period.
- Preprocessing: Data which is collected should be preprocessed redundant data, inconsistent should be taken care.
- Attribute Selection: Important Features must be extracted.
- Classification Algorithm: An appropriate and efficient algorithm should be employed.
- Result: prediction or recommendation can be provided to the farmers based on the results obtained.

E. Crop Replication Models

- *Background*

Crop Simulation Models (CSM) are computerized representations of crop growth, development and yield, simulated through mathematical equations as functions of soil conditions, weather and management practices (Hogeboom et al., 2004). The strength of the CSM is in their ability to extrapolate the temporal patterns of crop growth and yield beyond a single experimental site. Crop Simulation Models (CSM) can be used to gain new scientific knowledge of crop physiological processes or to evaluate the impact of agronomic practices on farmers' incomes and environments. Crop models are only an approximation of the real world and many do not account for important factors such as weeds, diseases, insects, tillage and phosphorus (Jones et al., 2001). Nevertheless, CSM have played important roles in the interpretation of agronomic results, and their application as decision support systems for farmers is increasing. Models range from simple to complex. The purpose a crop model is to be used determines to a large extent the complexity of a model. The crop growth system in general is more stochastic than deterministic because many parts of the agroecosystem are heterogeneous. However, to date, crop models using a stochastic approach have not been developed to a level of usefulness in decision making except in cases where year–to–year variations in weather are accounted for using deterministic models. Deterministic crop models can be classified into three basic types: statistical, Automatous, and functional

- *Statistical Models*

The first models used for large-scale yield simulations were statistical. Average yields from large areas (counties or crop reporting districts) and for many years were regressed on time to reveal a general trend in crop yields

(Thompson, 1969; Gage and Saffir 2011). The trend for the past several decades has been usually been upward and accounts for technological advancements 3 in genetic and management s especially the increased use of fertilizers. Deviations from the trend have been correlated with regionally averaged monthly weather data for each year. Thompson (1986) used a statistical type model to determine the impact of climate change and weather variability on corn production in five Midwestern states in the USA. He found preseason precipitation (September – June), June temperature, and temperature and rainfall in July and August to be closely correlated with corn yield variations from the trend. Recently Gage and Saffir (2011) incorporated climate effect with the use of the Crop Stress Index (CSI) into the regional yield trend. When simulating yield using statistical models, the effects of changes in agricultural technology must be subjectively extrapolated into time when the mix of the technology is unknown for that period.

- *Automatous Models*

Automatous models attempt to use fundamental mechanisms of plant and soil processes to simulate specific outcomes. Soon after computers became available for science, Automatous models were developed to simulate photosynthetic processes such as light interception, uptake of carbon dioxide (CO2), respiration and production of biomass partitioning biomass into various plant organs, and loss of CO2 during respiration. In the soil system, the Automatous approach was used to simulate the dynamics of water in the soil water due to infiltration, evaporation, drainage, and root uptake. de Wit (1965) recommended to distinguish between two levels, the system level and the next lower (explanatory) process level. Crop models typically consider the processes of plant development, light interception, CO2 assimilation and respiration, and the partitioning of biomass to plant organs and their growth.
Uncertainty in some assumptions makes Automatous model outcomes less certain and 4 often makes them less useful to those outside of the model development group. Automatous models are seldom used for problem solving purposes; rather, they are often used for academic purposes to gain a better understanding of specific processes and interactions. This and the increasing complexity of problems to be addressed have resulted in the development of more complex models. Automatous models usually describe instantaneous rates of plant processes that change rapidly over short time scales. For example, photosynthetic and transpiration processes change rapidly during the day as the radiation and temperature conditions change.

- *Functional Models*

Functional models use simplified approaches to simulate complex processes. In some cases, Automatous models can provide useful information that can be simplified into empirical functions for models. For example, many functional models use daily solar radiation as the amount of energy available for photosynthesis. The energy intercepted by the crop is approximated using feedback information from the plant leaf area index to approximate the biomass production using a simple concept of radiation use efficiency, e.g. the biomass produced per unit of radiation intercepted. Although this type function is much simpler than the more complicated ones, it usually produces reasonable results when compared to field measurements (Ritchie, 1980). Evapotranspiration is also simulated using only daily weather inputs by incorporating similar concepts to those used for biomass production simulation. In fact, the functional Penman equation for simulating potential evapotranspiration was being used two decades before computers were available for modeling and continues to be used, with some modifications, in crop models.
Functional models usually use simplified equations and logic to partition the simulated biomass into various organs in the plant, ultimately resulting in total biomass and economic yield. Evapotranspiration is used in a water soil balance equation to approximate when deficits or excesses in soil water or nutrients will impact potential biomass production and evapotranspiration. Functional models practically always use capacity concepts to describe the amount of water available to plants as compared to using instantaneous rate concepts from soil physics. A lower and upper limit of water capacity is defined as inputs, and water inputs and outflows in the soil provide the feedback to determine water availability to plants.

Table 1. Minimum data set needed to operate a crop replication model.

Input for Crop Replication Model
1. Site description
Latitude and longitude, elevation, average annual temperature
Slope and aspects of the site
2. Weather
Daily global soil radiation, daily maximum and minimum temperature, daily rainfall.
3. Soil
Soil type, soil depth (divided by n layers), soil texture, soil organic carbon, bulk density, soil nitrogen, pH
4. Initial condition of the system
Previous crop, residues left on the soil (if any), initial soil water and soil nitrogen
5. Crop and field management
Cultivar name and type, planting date and type, row space, plants per square meter, irrigation/nitrogen amount, method, dates of irrigation/fertilization, fertilizer type

TABLE I. Different approaches used for crop yield prediction

Algorithm /models used	**Crop type**	**Accuracy/ Recommendation**
Multiple Linear Regression [3]	Rice yield	90%-95%
Decision tree analysis and ID3[4]	Soybean	The rules formed from the decision tree is helpful in predicting the conditions

		responsible for the high or low soybean crop productivity under given climatic
Support Vector Regression model [5]	For any crop	The results show that support vector regression can serve as a better reference models for yield prediction. It is computationally less demanding.
Three models used APAR, SEBAL, Carnegie Institution Stanford model [6]	Wheat, rice , sugarcane, cotton	Successful for wheat , rice, sugarcane but not successful for cotton. The proposed technology can significantly contribute to quantitatively describing yield variations across the Indus Basin
Neural Networks [7]	Corn yield	95%
C4.5 algorithm and decision tree [8]	Soybean, paddy, maize	For soybean=87% For paddy=85% For maize=76%
Harmonic Analysis of NDVI Time-Series algorithm [9]	sugarcane	86.5%
Gaussian Processes [10]	Wheat yield	Wheat yield is expected to increase with an increase in temperature but there can be an increasing under estimation error in predicting the wheat yields
Relational cluster Bee Hive algorithm [11]	Any crop	This crop yield prediction model (CRY) performs 12% better than cluster & Regression Tree algorithm
K-Means algorithm for clustering And for classification Linear Regression, K-NN, ANN model [11]	Wheat, potato	90% - 95%
J48, LAD Tree [11]	Rice	100%
K-Nearest Neighbor (KNN) and Naive Bayes (NB) [11]	Any crop	classification of soil into low, medium and high categories are done

III. FUTURE WORK

It can be observed from table 1, there are still many challenges in this research area. To meet such challenges a more careful study has to be carried out. We should explore for a robust and novel classifier to improve the performance of the prediction system. We should be able to suggest based on the nutrient contents of the soil which crop is most suitable for a farmer's land. In addition, we should be suggesting water tolerant seed variety for sowing so that in case of flood and drought the crop could withstand the damage. Sensors can be used to measure the moisture and the nutrients in the soil, this information can also be used to guide the farmers

IV. CONCLUSION

According to our study, there is still scope for the improvement in result. During the study which we have carried out it is observed that the algorithm which is used by most of the authors do not uses a unified approach where in all the factors affecting the crop yield can be utilized simultaneously for predicting the crop yield. There is still further scope of improvement as the dataset which is considered is small in some cases. Therefore, the result can be improved by using a large dataset.

REFERENCES

[1] Maria Rossana C.de Leon, Eugene Rex L. Jalao, "A prediction model framework for crop yield prediction," Asia Pacific Industrial Engineering and Management System.

[2] Ramesh A. Medar, Vijay S. Raj purohit, "A survey on Data Mining Techniques for crop yield prediction", International Journal of Advance Research in Computer Science and Management Studies. Volume2, Issue 9, Sept 2014.

[3] D Ramesh, B Vishnu Vardhan, "Region specific crop yield Analysis: A data Mining approach ", UACEE International Journal of Advances in computer Science and its Applications-IJCSIA volume 3: issue 2.

[4] S. Veenadhari, Dr Bharat Mishra, Dr C D Singh, "Soybean productivity modeling using Decision Tree Algorithms," International Journal of Computer Applications Vol 27-No7 Aug 2011.

[5] George RuB, "Data Mining of Agricultural Yield Data: A comparison of Regression models.

[6] Wim G.M. Bastiaanssen, Samia Ali, "A new crop yield forecasting model based on satellite measurements applied across the Indus Basin, Pakistan," Agriculture, Ecosystems and Environment 94 (2003) 321–340.

[7] Sudhanshu Sekhar Panda, Daniel P. Ames, and Suranjan Panigrahi, "Application of Vegetation Indices for Agricultural

Crop Yield Prediction Using Neural Network Techniques", Remote Sensing 2010, 2, 673-696; doi:10.3390/rs2030673
[8] S.Veenadhari,Dr. Bharat Misra ,Dr. CD Singh," Machine Learning Approach for forecasting crop yield based on climatic parameters", 978-1- 4799-2352-6/14/$31.00 ©2014 IEEE
[9] Jefferson Lobato Fernandes; Jansle Vieira Rocha," Sugarcane yield estimates using time series analysis of spot vegetation images", Rubens Augusto Camargo Lamparelli, Sci. Agric. (Piracicaba, Braz.), v.68, n.2, p.139-146, March/April 2011
[10] Yunous Vagh, Jitian Xiao," Mining temperature profile data for shire-level crop yield prediction", 978-1-4673-1487-9/12/$31.00 ©2012 IEEE, Proceedings of the 2012 International Conference on Machine Learning and Cybernetics, Xian, 15-17 July 2012
[11] M. Gunasundari, Dr. T Arunkumar, Ms. R. Hemavathy," CRY – An improved Crop Yield Prediction model using Bee Hive Clustering.