Application of Data Mining Techniques for Crop Productivity Prediction

Zekarias Diriba
zekariaskifle@gmail.com

Berhanu Borena
PhD Candidate, Addis Ababa University, Ethiopia
berhanuborena@gmail.com

Abstract

Agriculture has been and is the main stay of Ethiopia's economy. It contributes to 45% of the GDP and engages more than 82% of the population. It is largely practiced by small holders and often vulnerable for environmental and weather shocks. As a result, the country's productivity is low by international standard. The low productivity doesn't come randomly. Rather it is associated with different factors. Currently crop productivity prediction is done using statistics. The current statistical analysis is inefficient and has no validation mechanism. In this paper, data mining is used for crop productivity prediction.

This study assesses predictive data mining applications that can be applied on Ethiopian agricultural crop productivity by focusing on small holder farmers. This study uses data from the Ethiopian Economic Association (EEA). The data mining process in this paper follows the CRISP. For the implementation of the final output, decision tree is used. There are three algorithms employed: namely J48, Random Forest, and REPTree. The model is represented in terms of IF THEN RULE. The tools used are Weka 3.7.7 and Excel 2007. The major finding of the result is that fertilizer use has the highest predictable power than the other factors, and the result of the final result can be used for policy makers, decision making process, and for further research to be done in the area.

Keywords: Decision Tree; Data Mining; Crop Productivity; Predicting Productivity

1. Introduction

Data mining is becoming increasingly important to discover knowledge patterns. It is a powerful technology that helps different knowledge workers to focus on most important data on their databases. Data mining tools predict future trends and behaviors, allowing businesses to make proactive and knowledge-driven decisions [1]. One of the areas where data mining could provide potential support is agriculture.

Currently, to predict crop productivity, statistical analysis is used. But statistical analysis is inefficient in showing all the knowledge representations. It has two major shortcomings. The first is it doesn't have sufficient validation mechanisms and the second is it takes time to run statistical programs [9]. In this study, predictive data mining methodology is used to predict crop productivity in Ethiopia.

The rest of the paper is organized in four sections. Section two presents the background information about crop productivity related issues. Section three covers related work. Section four covers the research method. Section five describes the results and the findings. Finally, conclusion and discussion is presented in Section six.

2. Background

Ethiopia's crop agriculture is complex, involving substantial variation in crops grown across the country's different regions and ecologies [5]. There has been substantial growth in cereals, in terms of area cultivated, yields and production since 2000, but yields are low by international standards and overall production is highly susceptible to weather shocks, particularly droughts [8]. Thus, both raising production levels and reducing its variability are essential aspects of improving food security in Ethiopia, both to help ensure adequate food availability, as well as to increase household incomes.

3. Related Work

The literature covers two components. The first part introduces data mining application in crop productivity and the second part presents a review of related works in the area of crop prediction.

The first literature covered is prediction of cotton growth and development [3]. There are 500 instances used with 24 attributes. The research uses quality attributes in the class label as Good, Average, and Bad. And there are 18 classifiers used including J48, Random Forest, REPTree, and Naïve Bayes. Naïve Bayes is found to be the most effective.

Another study predicts soybean productivity using decision tree algorithm [4]. It uses a spatial-temporal data input as factors for predicting crop yield. It assesses the relationship between metrological parameters and crop performance. The input attributes used are average rainfall, average evaporation, average maximum temperature, average maximum relative humidity, and average soybean yield. The method of classification is High Yield production and Low Yield production. Id3 algorithm is used and the generated tree result is modeled using IF THEN RULE for knowledge discovery.

A multi-temporal image is taken in northern Taiwan to interpret the rice paddy distribution on the year 2000 using the Bayesian classifier [6]. The data input used is the remote sense data as an input for spatial temporal imaginaries. The method to achieve the final objective of the study is by classifying as paddy and non-paddy distribution of the crop production.

Another study depicts the relationship between geospatial data and agricultural out production applied using decision tree [7]. The data is collected in Andhra Pradesh in India from 1991/1992 up to 2000/2001 in which the use of land utilization like forest coverage and uncultivated land amount in the state is used as input data. IF THEN RULE is used for knowledge discovery. It compares the result whether the demand and supply are demanded of water, fertilizer and pesticide usage and which factor has the highest impact for final crop productivity.

The second part of the literature focuses on the decision tree. Decision tree is preferred to be less complex in which less complexity can be expressed in the total number of nodes, total number of leaves, tree depth, and number of attributes.

Decision tree in general is a two-step process. One is splitting criteria and the second is pruning technique. Decision tree splits the attributes based on different Univariate criteria [2]. Univariate means that an internal node is split according to the value of a single attribute. There are different Univarite criteria covered like Information Gain, Gini Index, and Gain Ratio. Each of them has different theoretical approach as a splitting mechanism. The second step of decision tree is pruning technique. Employing tightly stopping criteria tends to create small and under-fitted decision trees [2]. On the other hand, using loosely stopping criteria tends to generate large decision trees that are over-fitted to the training set. Pruning methodology solves this dilemma. A node is pruned if this operation improves a certain criteria. The most popular pruning techniques are Cost-Complexity Pruning, Reduced-Error Pruning, Minimum-Error Pruning (MEP), Pessimistic Pruning, Error-Based Pruning (EBP), Optimal Pruning, and Minimum Description Length Pruning.

Finally and in addition to the above major literature areas, the four performance metrics are covered that are ROC, F-measure, precision, and recall.

4. Research Method

The real world data can be characterized by missing values, inconsistency, and extreme values (outliers). For this purpose, the data must be processed before it is experimented. Preprocessing is important in order to meet the final objective effectively. The data has no problem with consistency rather with incompleteness like missing values and outliers and data must be aggregated.

The major tasks done are making the data to be complete for final modeling. The major tasks done on data preprocessing are data cleaning, outlier detection and removal, replacement of missing values, and aggregation the data.

4.1 Data Understanding

Before all the data preprocessing, data understanding is needed. In order to accomplish the study, the data is collected from Ethiopian Economic Association (EEA). The data is originally collected for the statistical analysis research of land tenure system in Ethiopia. The data is constructed in 80 questionnaires and these questionnaires are represented in terms of codes and located in different locations that must be merged. From these data, four groups of factors are identified: Environmental factors (Crop rotation, Tree planting, Terracing), Land condition, Agricultural technology input (Fertilizer use i.e., UREA and DAP), and marketing (market distance and credit service).

4.2 Data Cleaning

The tasks done on data preprocessing part are missing handling values, outlier detection, and removal.

Missing values handling: Missing values are handled by replacing with question mark (?).Those with too much missing values, the attributes are removed like education status 41% and pesticide usage with 86%.

Outlier detection and removal: Outliers are observations that deviate a lot from the data pattern. In order to detect the data deviation based algorithm is used. Three times the deviation the distance from the mean value is used to detect the outlier.

3SD Method= Mean +- 3SD

After the outlier is detected it is filter using Excel and these filter value is replaced with a mean value.

4.3 Attribute Selection

Out of the total 90 attributes, 13 attributes are selected. Attributes that are removed are not important to meet the final objectives and with too much missing values. The selected attributes are illustrated in Table 1.

Table 1: The selected attributes

Attribute Name	*Type*	*Description*
Fertilizer Use	T	Fertilizer use amount in terms of high and low (above the average amount and below the average amount respectively)
Fertilizer per hectare	N	Total fertilizer used per cultivated land
Credit service	T	Credit service(Yes or No)
Crop Rotation	T	Crop rotation(Yes or No)
Market distance	N	The nearest available market for the farmer
Soil fertility	T	Soil fertility(Yes or No)
Mixed Cropping	T	Mixed cropping(Yes or No)
Land Management	T	Land Management (Yes or No)
Terracing	T	Terracing (Yes or No)
Tree planting	T	Tree planting (Yes or No)
Fallowing	T	Fallowing (Yes or No)
Land	N	Total cultivated amount (hectare)
Productivity label (Total quintal per hectare)	N	The productivity label based on the amount produced per hectare

4.4 Aggregating and merging files

It the most time consuming tasks from all the data preprocessing tasks, since all the data are located in different files, they must be merged. It uses question Id used for each household to merge the recorded data. Different files with the same question Id represent the same house collected data record.

4.5 Assigning Class Label

The target variable for class labeling is crop productivity per hectare. Because it is an outcome of the result the input factors, it is used as a class labeling. The crop productivity prediction was classified as High productivity and Low Productivity. High productivity is assigned above the average value of Ethiopia's 16.8 quintal per hectare and below the average value of 16.8 quintal per hectare [8].

5. Result and Findings

The data is experimented using N-fold cross validation for testing and training data. The dependent variable is Total quintal per hectare and all others are independent variables. Two experiments are done. The first one uses J48 and the second one REPTree. No experiment is presented for Random Forest because there is no result generated.

5.1 Experiment One

To shorten the J48 generated tree result, a minimum of 100 instances are used. It is further represented in IF THEN RULE. From all variables, Fertilizer is the highest classifier result. According to the J48 algorithm, 94.22 % of them resulted in high crop productive and are found to be high crop fertilizer use.

5.2 Experiment Two

The second experiment is tested using REPTree. To generate a result that can be explained very easily, the research uses tree depth of five. IF THEN RULE is used to generate the final outcome of the experiment. According to the REPTree algorithm result 55.4 % of them high crop productive are found to be high crop fertilizer use.

5.3 Information Gain Value

Attributes are ranked based on their information gain value. InfoGainAttributeEval is used as evaluator and Ranker is used for searching mechanism. The ranked values are the following.

1. Fertilizer Use
2. Land
3. Credit Service
4. Crop Rotation
5. Market Distance
6. Soil Fertility
7. Mixed Cropping
8. Land Management
9. Terracing
10. Tree Planting
11. Fallowing

5.4 Algorithm Evaluation

Table 2: Evaluation of the three algorithms

	J48	*REPTree*	*RandomForest*
Total records	8540	8540	8540
Correctly classified in No	7105	7080	7122
Correctly classified in Percentage	83.12	82.904	83.39

From Table 3, we can see a slight difference in the performance between the three algorithms. The above result in general shows how much of the result is correctly classified. REPTree performed higher; next is RandomForest and J48 is relatively the least performing algorithm.

5.5 Performance Evaluation Metrics

Table 3: Evaluation metrics of the three algorithms

Measure	*J48*	*REPTree*	*Random Forest*
ROC	0.846	0.989	0.867
Precision	0.825	0.938	0.824
Recall	0.832	0.936	0.829
F-Measure	0.828	0.932	0.827

As shown in Table 4, the three algorithms have shown to be very closely efficient with the efficiency evaluations. But there is a slight difference. ROC curve shows that REPTree and J48 are more efficient. Precision shows that REPTree and Random Forest are more efficient. Recall shows that REPTree and Random Forest are more efficient. F-Measure shows that REPTree and Random Forest are more efficient.

In the result, other than ROC, J48 is slightly less efficient. It is because the data is more skewed rather than discreet and more dispersed.

Table 4: Resampling of the three algorithm results

Percentage Used	*J48*	*REPTree*	*Random Forest*
20 %	83.37	82.79	85.59
40 %	83.19	83.39	82.90
60 %	85.61	85.42	88.52
80 %	89.44	88.01	93.92

Resampling of the data is used to know how much accuracy it has on different percentages of the data. In this study 20, 40, 60, and 80 percent of the data is sampled just to check the efficiency similarity as shown in Table 4.

6. Discussion and Conclusion

The overall objective of this study is to predict crop productivity. It is believed to support policy and decision makers in order to make their policy making process scientifically supported. It is believed that it will help small scale farmers to increase the outcome of their crop production.

This study recommends for policy makers to make proactive decisions in identifying which factors are the most important to increase productivity. This study shows not only the use of fertilizer but the amount of fertilizer matters in productivity. Therefore, the experts should deploy this outcome of the data mining result in providing sufficient fertilizer amount.

From the result, we can conclude three major things. The first one is out of all attributes used, fertilizer use has the highest predictive power. The second one is, out of the three algorithms tested, J48 has shown more predictive power. Of course all of the three algorithms have shown almost the same efficiency. The third conclusion is that the data may not have efficient predictable power as only one year data and only 8540 rural household records are collected out of millions of records.

References

[1] U. M. Fayyad, G. P. Shapiro, P. Smyth, and R. Uthusrsay, “Advances in Knowledge Discovery and Data Mining”, 3rd ed., 1996.

[2] Lior Rokach and Oded Maimon, “Top-Down Induction of Decision Trees Classifiers - A Survey”, 2005.

[3] P. Revathi and M. Hemalatha, “Efficient Classification Mining Approach for Agriculture”, 2011.

[4] S. Veenadhari and B. Mishra, “Soybean Productivity Modelling using Decision Tree Algorithms”, 2011.

[5] J. W. Mellor and P. Dorosh, “Agriculture and the Economic Transformation of Ethiopia”, 2010.

[6] Chi-Chung Lau and Kuo-Hsin Hsiao, “Bayesian Classification for Rice Paddy Interpretation”, 2011.

[7] Ahsan Abdullah, Stephen Brobst, and Ijaz Pervaiz, “Learning Dynamics of Pesticide Abuse through Data Mining”, 2004.

[8] Crop Production in Ethiopia: Regional Patterns and Trends.

[9] Guidelines for the Validation And Verification of Quantitative and Qualitative Test Methods, National Association of Testing Authorities, 2012.