A DATA MINING APPROACH TO CROP YIELD PREDICTION

Joshua Ward[1], Noah Clark[2], Isabella Adams[3], and Samuel Scott*[3]
[1] Department of Computer Science, University of British Columbia, Canada
[2] Department of Physics, University of Cambridge, UK
[3] School of Civil Engineering, University of Queensland, Australia

ABSTRACT

We all know agriculture is the most important factor which influence the economy of India and it also offers employment to 50% population of India. People of India are practicing agriculture for many years and the result were never satisfying due to many factors that effect the crop. Day by day environment is changing and is not stable at various places. It is very important for the farmer to cultivate their farm in good climatic conditions ,under such conditions they need technology that predict the environment. In such cases data mining is the apt technology for prediction. Data mining contains various prediction algorithms like id3,cart ,c4.5,random forest algorithm. In this Project we are using Id3 and cart algorithms as a prediction techniques and it is possible obtain the information from the prediction algorithms which helps farmers to cultivate the appropriate crop.

KEYWORDS: Datamining, crop yield prediction,ID3 and Cart algorithm.

1. INTRODUCTION

Crop yield prediction is an important area of research which helps in ensuring food security all around the world. India has rich soil which is very ideal for cultivation and it is one of the top rice producing countries in the world. In order to take full advantage of the soil and sub-tropical climate of India, farmers need to know exactly when to plant crop seeds. The entire economy also depends on the produce from harvesting annually.

Different districts in India have varying climates and so it is very important to consider environmental factors of these separate areas. This will help to choose the best districts for cultivation of different type of crops. Rainfall also varies from district to district and this has a huge impact on farming because while too little or too much rain can kill crops, the proper amount of rain leads to an ideal crop yield. With rainfall comes humidity and since rainfall varies from district to district so does humidity. Humidity causes changes in the level of water that can be absorbed by atmosphere which can cause crops to remain too wet or too dry and so to get proper yield, a district with an ideal average annual rainfall and humidity is required.

Finally, the most important part of farming has to be considered, pesticides. Without pesticides crops would die significantly more due to insects and other pests leading to a sudden drop in yield. Too much pesticides may affect the crop on its own while too little may not get rid of pests. So, the amount of pesticides required by crops is a very important parameter.
In our research, we have considered the effects of environmental(weather), biotic(pH, soil salinity) and area of production as factors towards crop production in Bangladesh. Taking these factors into consideration as datasets for various districts, we applied clustering techniques to divide regions; and then we apply suitable classification techniques to obtain crop yield predictions.

2. RELATED WORK

Ramesh and Vardhan [1]deal with the challenge of predicting the yield of various crops. One approach to this problem is to employ data mining techniques. In this paper, different types of data mining methods were applied

and then evaluated on the datasets we prepared. In [2], Diepeveen and Armstrong discuss about various crop related data that is supplied to farmers to make better decisions to enhance yield and profits. While this may give the advantage of a particular crop species over others, the data is generalized and may not apply to others. There are data mining application that can process the data and improve the quality and reliability of this dataset for different farming situations. The challenge is identifying key attributes that affect crop yield, such as geographic location, soil type, seasonal conditions, nutrition, grain yield and quality, sowing and harvest data and tolerance to environmental stress. In this paper data mining techniques were used to help growers find the combination of traits required to identify high performance species. Several techniques were used over different geographical locations. In [6] Murynin et al. study the dependency between the prediction and the accuracy of the forecast. The linear model is selected as a basic approach of yield prediction. Then, the model is extended with non-linear attributes in order to improve the accuracy of the prediction. The extensions take into

consideration long-term technological advances in agricultural productivity as well as regional variations in yields. The accuracy of the model has been estimated based on the time period between the moment of the forecast formation and the time of harvest.

3. DATA SET

The dataset used in this research has been collected from IARI (Indian Agricultural Research Institute). All the data were in pdf format which were converted to rtf format using miscellaneous tools and tricks. A lot of preprocessing was required to handle missing values, noise and outliers. From the dataset, we have preprocessed and selected only the attributes which are important for our research: rainfall, maximum and minimum temperature, humidity, irrigated area for all districts; and cultivated area for every crop considered according to the districts. One further environmental attribute: sunshine and two further biotic attributes – soil salinity and soil pH were considered for our research. These data were collected from the Bangladesh Agricultural Research Council (BARC) website [9]. After the necessary formatting and preprocessing of the datasets, the finalized version of our data contained a total of 15 districts for the time periods of 2009-10 and 2010-11. The crop yields were selected for the following crops which have been considered for our project:

- Rice- AMON
- Rice
- SugarCane
- Wheat

4. INPUT VARIABLES

From the vast initial dataset, we selected a limited number of important input variables which have the highest contribution to agricultural produce.

a) The environmental variables:

i) Rainfall: The average yearly rainfall was considered by calculating average from the monthly rainfall (mm) of each district. Usually, the year that contains the highest average rainfall should provide for maximum crop yield in that year. ii) Humidity: Similar to the way we collected the rainfall data, we also calculated and obtained the average yearly humidity for each district in percentage. iii) Max Temperature: variation in temperature through the year puts a great impact in that year's crop production. Hence we consider both the maximum as well as the minimum temperature in our research. iv) Min Temperature: The average yearly minimum temperature (considered in Celsius). v) Average Sunshine: The amount of sunshine received on areas each year greatly effects the production of green crops as it directly affects the photo-synthesis process in plants. This attribute was considered in hours as a yearly average for each district.

b) The biotic input attributes:

i) Max pH: Maximum pH of a district's soil. pH is a scale attribute for farmers to keep track for how acidic the soil is. This scaled is defined by a value of 7, where soil pH above 7 meaning alkaline and below 7 meaning acidic. Crop production is highly affected by the variations of pH in soil. ii) Min pH: Minimum pH of a district's soil. iii) Soil Salinity: Taken as MMHOS/cm, the ranges were (<2), (2-4), (4-8) and (8-15). Soil salinity defines the amount or content of salt in soil. Soil salt content is increased by the process of salinization. Too high soil salinity can cause a detrimental effect towards crop production and yield. We calculated total areas (in hectares) under different salinity ranges for each of the 15 districts.

5. EXISTING SYSTEM

With the evolution of the algorithms in data mining. The prediction process is changing in terms of speed with the use of data mining techniques and new algorithms. But the existing systems lack in terms of speed and efficiency due to implementation of techniques with high time complexity and implementation of primitive algorithms. Even if a particular website tries its best to grab any customers, there is a huge competition from the market. Website owners are thus, unable to understand the user's personal needs and as a result are failing to meet their demands.

6. PROPOSED SYSTEM

The method of our research is initially divided into two major parts: Front End, Classification.

A. Front End:In this we design a graphical user interface (GUI) in order to browse and upload the dataset and also for to show the decision tree.

B. Prediction of crop yields using classification techniques:

In our research, we determined prediction results for yields of selected crops for the selected districts.

The predictions results were obtained according to the selected input attributes using appropriate classification.

The following classification were used to obtain the crop yield prediction results:

Algorithm 1: ID3 Algorithm:
Decision tree algorithms transform raw data to rule based decision making trees. Herein, ID3 is one of the most common decision tree algorithm. Firstly, It was introduced in 1986 and it is acronym of **Iterative Dichotomiser**. dichotomisation means dividing into two completely opposite things. That's why, the algorithm iteratively divides attributes into two groups which are the most dominant attribute and others to construct a tree. Then, it calculates the entropy and information gains of each atrribute. In this way, the most dominant attribute can be founded. After then, the most dominant one is put on the tree as decision node. Thereafter, entropy and gain scores would be calculated again among the other attributes. Thus, the next most dominant attribute is found. Finally, this procedure continues until reaching a decision for that branch. That's why, it is called Iterative Dichotomiser. So, we'll mention the algorithm step by step in this post.
We can summarize the ID3 algorithm as illustrated below
Entropy(S) = ∑ – p(I) . log2p(I)
Gain(S, A) = Entropy(S) – ∑ [p(S|A) . Entropy(S|A)]

Algorithm 2: CART Algorithm
Classification and Regression Trees implies the technique of recursively separating the observations in branches to build a tree to improve the prediction accuracy and predicts continuous dependent variables and categorical predictor variables. The CART algorithm was popularized by Breiman et al. (Breiman, Friedman, Olshen, & Stone, 1984; see also Ripley, 1996). Although after many investigations and enhancements by the researchers less research is done on enhancing CART performance in disease diagnosis especially in diagnosis of heart disease. In this paper, a method which is existing is applied to detect heart disease to obtain the result.CART utilizes Gini index to scale the impurity of a section or collection of training tuples. It is capable to handle high dimensional categorical data. Decision Trees can also handle continuous data but they must be turned into categorical data. This simplicity is useful not only for purposes of rapid classification of new observations (it is easier to evaluate one or two logical conditions, than to compute classification marks for possible groups, or predicted values, basing on all predictors and using possibly some complex nonlinear model equations),can also result a simpler "model" to explain why observations are classified or predicted in a particular manner. The final outputs of using tree methods for classification or regression can be summarized in a sequence of logical if-then conditions. Therefore, there is no implicit assumption that the underneath relationships between the predictor variables and the dependent variable are linear,follow specific non-linear link function, or that they are even

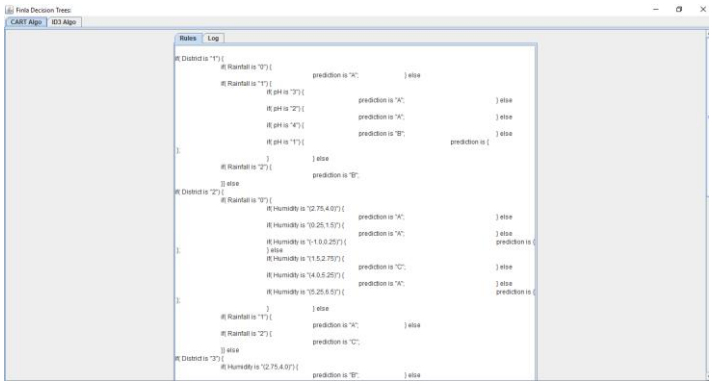
7. RESULTS

A. Clustering result analysis: We applied clustering to find if any strong correlation exists between crop yield and different attributes (i.e. weather attributes, soil PH and soil salinity attributes and area cultivated attribute). Although there are some similarity between the clusters obtained using different attributes with the clusters obtained according to crop yields, we did not find any exact or strong

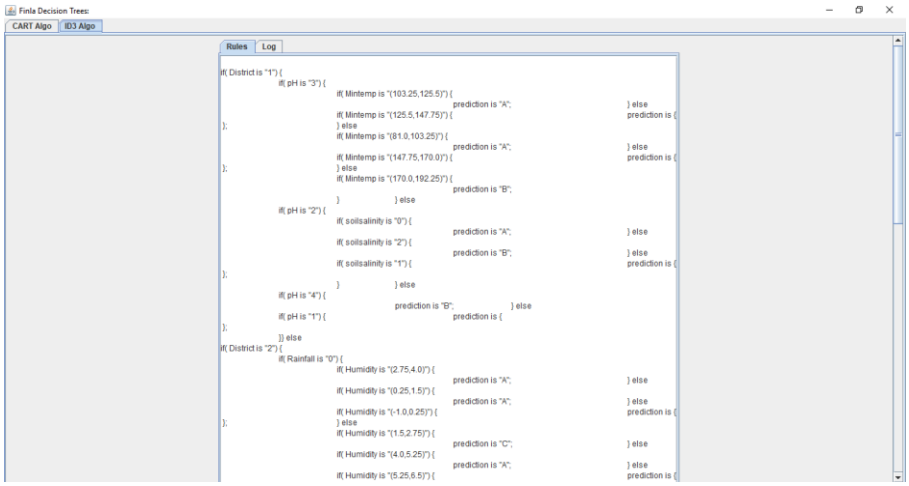
correlation between crop yield with weather/ soil/ cultivate area attribute. This might be due to the fact that there were some missing values in our data sets and the size of our data set was comparatively smaller than required to find any strong correlation B. Analysis of prediction results from different models:

B. From all the results obtained it is clearly seen that the accuracy lies within the range of 90 to 95 percent. Each of the techniques used gives a prediction with a slightly varying accuracy. However, due to the small training set, the prediction was not as accurate as expected and sometimes anomalies were experienced. For example, from FIG. 2 to FIG. 5, in some cases, if the actual yield was 0 (zero), our models sometimes erroneously predicted some nonzero value for the predicted yield. If our training dataset were large enough (containing all the data about all 64 districts), avoiding this problem would've been possible.

Cart rules



ID3 rules



8. CONCLUSION

In our research we have found that the accurate prediction of different species of crop yields across several districts could help a lot of farmers and others alike. A farmer could plant different crops in different districts based on simple predictions made by this research and if that does take into effect, each and every farmer would get a chance at increasing their profits and increasing the country's overall produce. Also, using a better dataset for this research will lead to even better predictions and recommendations as the recommendation engine is basing its decision based on the predictions.

REFERENCES

[1] D Ramesh , B Vishnu Vardhan. "Data Mining Techniques and Applications to Agricultural Yield Data". International Journal of Advanced Research in Computer and Communication Engineering Vol. 2, Issue 9, September 2013,pp.3477-3480.

[2] D. Diepeveen and L. Armstrong, "Identifying key crop performance traits using data mining" World Conference on Agriculture, Information and IT, 2008.

[3] Mohammad Motiur Rahman, Naheena Haq and Rashedur M Rahman "Comparative Study of Forecasting Models on Clustered Region of Bangladesh to Predict Rice Yield", 17th. IEEE International Conference on Computer and Information Technology (ICCIT), Dhaka, 2014.

[4] http://books.irri.org/0471097608_content.pdf

[5] http://www.assignmentpoint.com/science/zoology/agriculture-sector-ofbangladesh.html

[6] Alexander Murynin, Konstantin Gorokhovskiy and Vladimir Ignatie "Efficiency of crop yield forecasting depending on the moment of prediction based on large remote sensing data set" retrieved from http://worldcomp-proceedings.com/proc/p2013/DMI8036.pdf

[7] Ye, Nong; Data Mining: Theories, Algorithms, and Examples, CRC Press, 2013.