Soybean Productivity Modelling using Decision Tree Algorithms

S. Veenadhari
Research Scholar
MGCGV, Chitrakoot

Dr. Bharat Mishra
Associate Professor
MGCGV, Chitrakoot

Dr.CD Singh
Senior Scientist
CIAE, Bhopal

ABSTRACT
Data mining applications in agriculture is a relatively new approach for forecasting / predicting of agricultural crop/animal management. In the present study an attempt has been made to study the influence of climatic parameters on soybean productivity using decision tree induction technique. The findings of Decision tree were framed into different rules for better understanding by the end users. The study findings will help the researchers, policy makers and farmers in predicting/forecasting the crop yield in advance for market dynamics.

Keywords— Decision tree, crop productivity, ID3 algorithm, climatic factors

1. INTRODUCTION

Agriculture and allied activities constitute the single largest component of India's gross domestic product, contributing nearly 25% of the total. Agriculture in India continues to be fundamentally dependent on the weather conditions which are erratic and influence the crop productivity. Agricultural productivity is sensitive to two broad classes of climate-induced effects [1] direct effects from changes in temperature, precipitation, or carbon dioxide concentrations, and [2, 3] indirect effects through changes in soil moisture and the distribution and frequency of infestation by pests and diseases. Statistical regression and simulation models have been developed in assessing the relationship between the meteorological parameters and crop performance.

Soybean is one of the most predominant crop cultivated in the state of Madhya Pradesh, India. Over the past two decades the productivity of the crop has been in declining trend despite the area under the crop is increasing. Using the long term meteorological data it is now possible to predict the influence of different meteorological parameters on the crop yield using decision tree induction approach.

The major agricultural inputs and their effect on crop yield and also cost of cultivation affected by different inputs in selected States of India[15]. In this study regression analysis was used in predicting the input interaction on the crop yield.

A model for real time assessment of the direction and quantum of variability in wheat yields was developed [9]. A simple technology trend model in conjunction with crop simulation model (CERES-Wheat in DSSAT environment) was used for early wheat yield prediction at six locations representing the six major wheat-growing states, which contribute about 93% of national wheat production. A three-step approach, viz. (a) prediction of technological trend-based yields, (b) quantification of weather-induced yield variability using Crop Simulation Model (CSM), and (c) final yield prediction combining the previous two steps (a) and (b), was applied. A simulation model when run on a common set of soil properties, genetic coefficients and agronomic practices, is supposed to capture inter-annual yield variability due to year-to year varying weather conditions. Deviation in observed wheat yield from its technology trend and deviation in simulated wheat yield from its trend/ average showed positive relationship (r = 0.57, $P > 0.05$). An overall RMSE of 0.158 t ha^{-1} (5.619%) with R^2 0.97 was found against mean wheat yield of 2.815 t ha^{-1}. Real time weather data up to February and normal onward were used, for early wheat yield assessment at six locations.

A case study of interpreting paddy distributions of three counties on Northern Taiwan during two crop seasons on year 2000 using multi-temporal imageries together with cadastre GIS by Bayesian posterior probability classifier was studied [5]. In order to integrating Bayesian conditional probability, priori probabilities of paddy's attributes were estimated from photogrammetric interpretation results provided by the Food Bureau, and the spectrum reflectance from different growth stages was used. Due to the spatial heterogenous of paddy's distribution, classifier parameters were established individually on each map-quadrangle. Temporal change of NDVI from different growth stages pass through rice's life cycle has been measured and we find two-stage images make significant improvement on classification results. Results of the study help us to evaluate the accuracy of the classifier. Imagery classification results were compared with aerial photo's interpreting results for assessing accuracy. Overall accuracy of first crop of Tao-yuan, Hsin-chu, and Miao-li were 89.93% 92.83% 95.33% respectively. Bayesian classifier has advantages including easy-to-adjusted and easy-to-computed rules and comparative stable results when limited SPOT satellite imageries available. Bayesian method also provides results with probability that help the operator to assess the places having least confidence. These advantages allow us to suggest Bayesian method be used in paddy-area investigation in Taiwan.

A process model was developed [6] for analysing data, using Waikato Environment for Knowledge Analysis (WEKA) in the model. The domain model learned by the data mining algorithm can then be readily incorporated into a software application. This WEKA based analysis and application construction process was illustrated through a case study in the agricultural domain i.e., in mushroom grading.

Analytical exploration of vast amount of agricultural data can best be supported by an appropriate application. [2, 3] applied Data warehousing and Online Analytical Processing (OLAP) technologies for appropriate utility of agricultural data. A data warehouse provides a flexible yet efficient and reliable storage structure for vast amount of

data while OLAP techniques provide mechanisms for ad hoc and in depth analysis of this data. Traditional analytical tools and database techniques may not succeed here due their rigid nature. Techniques used in their work are equally applicable at any geographic location provided that related data is available.

Data characteristics which affect the performance of naive Bayes was found out using the approach Monte Carlo simulations. [12]. The impact of the distribution entropy on the classification error, showing that low-entropy feature distributions yield good performance of naive Bayes was analysed. They also demonstrated that naive Bayes works well for certain nearly functional feature dependencies, thus reaching its best performance in two opposite cases: completely independent features (as expected) and functionally dependent features (which is surprising). The accuracy of naïve Bayes is not directly correlated with the degree of feature dependencies measured as the class conditional mutual information between the features. Instead, a better predictor of naive Bayes accuracy is the amount of information about the class that is lost because of the independence assumption. In a study "Measuring the impact of climate change on Indian Agriculture, data pertaining to Indian agricultural, climatological, edaphic and geographical variable over a period of 30 years were collected and compiled in the form of data sets for analysis [14].

Possibilities were explored to build an alarming system based on the results of the application of data mining (DM) techniques in genetic evaluations of dairy cattle, in order to assess and assure data quality[13]. The technique used combined data mining using classification and decision-tree algorithms, Gaussian binned fitting functions, and hypothesis tests. Data were quarterly national genetic evaluations, computed between February 1999 and February 2003 in nine countries. Each evaluation run included 73,000–90,000 bull records complete with their genetic values and evaluation information. Milk production traits were considered. Data mining algorithms were applied separately for each country and evaluation run to search for associations across several dimensions, including bull origin, type of proof, age of bull, and number of daughters. Then, data in each node were fitted to the Gaussian function and the quality of the fit was measured, thus providing a measure of the quality of data. In order to evaluate and ultimately predict decision-tree models, the implemented architecture can compare the node probabilities between two models and decide on their similarity, using hypothesis tests for the standard deviation of their distribution. The key utility of this technique lays in its capacity to identify the exact node where anomalies occur, and to fire a focused alarm pointing to erroneous data.

Thematic information related to agriculture which has spatial attributes also was studied with an aim at discerning trends in agriculture production with reference to the availability of inputs [8]. The Predicted and Real vs. Counter graph illustrates how closely the Poly-Analyst prediction follows the actual value of the attribute over the range of the dataset. Applying the data mining techniques to agriculture the target for different food grains is achieved. The study demonstrated the scope for application of spatial mining tools for a utility study and analysis. The specific application of Poly-analyst gave a clear scope for evaluation and comparison of predicted and real values.

From the literature reviewed it is evident that the application of data mining techniques in the field of agriculture in Indian subcontinent is very limited. Some of the studies carried out in India statistical tools were used to develop regression equations to find out the various parameters influence. The limitation of the earlier studies were either they are carried out in a smaller area or that the influence of different input parameters can't be judged. Therefore, the present study was proposed to develop innovative applications of data mining techniques in predicting influence of agro-climatic factors on soybean crop production in Bhopal district.

2. METHODOLOGY

Decision tree induction technique [7] is adopted in the present study to develop innovative approaches to predict the influence of climatic parameters on the predominant crop (soybean) productivity of Bhopal district. A decision tree is a flow-chart-like tree structure, where each internal node denotes a test on an attribute, each branch represents an outcome of the test, and leaf nodes represent classes or class distributions. The top most node in a tree is the root node. In order to classify an unknown sample, the attribute values of the sample are tested against the decision tree. A path is traced from the root to a leaf node that holds the class prediction for that sample. Decision trees were then converted to classification rules using IF-THEN-ELSE.

Interactive Dichotomizer 3 (ID3) is one of the decision tree algorithm adopted [10,11] in this study which is information based method that depends on two assumptions.

Let C contain p objects of class P and n of class N. The assumptions are:

(1) Any correct decision tree for C will classify objects in the same proportion as their representation in C. An arbitrary object will be determined to belong to class P with probability p/(p + n) and to class N with probability n/(p + n).

(2) When a decision tree is used to classify an object, it returns a class. A decision tree can thus be regarded as a source of a message 'P' or 'N', with the expected information needed to generate this message given by

$$I(p,n) = -\frac{p}{p+n}\log_2\frac{p}{p+n} - \frac{n}{p+n}\log_2\frac{n}{p+n}$$

If attribute A with values $\{A_1, A_2, \ldots A_v\}$ is used for the root of the decision tree, it will partition C into $(C_1, C_2, \ldots C_V\}$ where C_i contains those objects in C that have value Ai of A. Let *Ci* contain pi objects of class P and *ni* of class N. The expected information required for the sub tree for *Ci* is $I(p_i, n_i)$. The expected information required for the tree with A as root is then obtained as the weighted average

$$E(A) = \sum_{i=1}^{v} \frac{p_i + n_i}{p+n} I(p_i, n_i)$$

Where the weight for the i^{th} branch is the proportion of the objects in C that belong to Ci.

The information gained by branching on A is therefore

$$gain(A) = I(p,n) - E(A)$$

A good rule of thumb would seem to be to choose that attribute to branch on which gains the most information.

Decision Tree Induction: The basic algorithm for decision tree induction is a greedy algorithm that constructs decision trees in a top-down recursive divide-and-conquer manner. The basic strategy for the algorithm is as follows:

- The tree starts as a single node representing the training sample
- If the samples are all of the same class, then the node becomes a leaf and is labeled with that class.
- Otherwise, the algorithm uses an entropy-based measure known as information gain as a heuristic for selecting the attribute that will best separate the samples into individual classes. This attribute becomes the test or decision attribute at the node. In this version of the algorithm, all attributes are categorical, that is, discrete-valued. Continuous-valued attributes must be discretized.
- A branch is created for each known value of the test attribute, and the samples are partitioned accordingly.
- The algorithm uses the same process recursively to form a decision tree for the samples at each partition. Once an attribute has occurred at a node, it need not be considered
- The recursive partitioning stops only when any one of the following condition is true:
  a) all samples for a given node belong to the same class
  b) there are no remaining attributes on which the samples may be further partitioned. In this case, majority voting is employed. This involves converting the given node into a leaf and labeling it with the class in majority among samples.
  c) there are no samples for the branch test-attribute = $_{ai}$. In this case, a leaf is created with the majority class in samples.

3. RESULTS AND DISCUSSION

The sample climatic data collected for the period 1984-2003 are presented in Table 1. Therefore, to establish the most influencing climatic parameter on yield of soybean crop produced in Bhopal district an ID3 algorithm was used. The Climatic factors of Bhopal district considered in the analysis are presented below:

Average Rainfall : 995 mm
Average Evaporation : 6.2 mm/day
Average Maximum Temperature : 31.1^{o}C
Average Maximum Relative Humidity : 66.2%
Average Soybean Yield : 845 kg/ha

Table 1: Agro-Climatic data in Bhopal district during 1984 – 2003.

Y	R	E	T	RH	Soybean Yield, q/ha
1984	844.7	4.6	22.5	54.2	801
1985	1208.7	6.4	29.5	75.3	765
1986	1367.3	6.2	30.6	63.2	588
1987	645.6	6.8	31.1	55.8	582
1988	876.4	6.5	31.9	59.6	891
1989	670.5	6.7	31.6	56.7	801
1990	1222	5.7	30.7	71.5	1018
1991	879.5	6.0	31.7	71.2	792
1992	767.1	6.7	32.1	68.7	869
1993	1087.3	6.2	31.7	73.5	893
1994	1241	5.2	31.3	74.8	804
1995	741.1	5.8	31.8	73.3	929
1996	1336.7	6.2	31.8	66.8	805
1997	1149.5	6.3	32.1	69.1	955
1998	1057.4	6.4	32.4	67.2	928
1999	1316.1	6.1	30.9	69.1	895
2000	773.7	6.5	32.3	62.3	860
2001	779.6	6.3	31.7	62.5	900
2002	792	7.1	32.9	60.8	758
2003	1154	6.0	31.3	68.9	1074
Ave.	995	6.2	31.1	66.2	845

Table 2: Details of attributes

Attribute	Explanation
Y	Year
R	Rainfall, mm
E	Evaporation, mm/day
T	Temperature, ^{o}C
RH	Relative Humidity, %

From average of each parameter was considered to classify the data into high, equal and low values. Thus classified data is presented in Table 3.

Table 3 Classification of Agro-climatic data based on average values

Y	R	E	T	RH	Soybean yield, q/ha
1984	L	L	L	L	L
1985	H	H	L	H	L
1986	H	E	L	L	L
1987	L	H	L	L	L
1988	L	H	H	L	H
1989	L	H	H	L	L
1990	H	L	L	H	H
1991	L	L	H	H	L
1992	L	H	H	H	H
1993	H	E	H	H	H
1994	H	L	H	H	L
1995	L	L	H	H	H
1996	H	E	H	H	H
1997	H	H	H	H	H
1998	H	H	H	H	H
1999	H	L	L	H	H
2000	L	H	H	L	H
2001	L	H	H	L	H
2002	L	H	H	L	L
2003	H	L	H	H	H

Classification by decision tree induction for parameters influence on soybean yield was worked using the equation I(p, n) explained in the methodology for high and low values of rainfall, evaporation, temperature and relative humidity individually by comparing the values with the average values of the Bhopal district. The entropy and gain values were also calculated for these parameters using equations described in the methodology.

From the induction tree analysis it was found that the influence of Maximum relative humidity has a maximum gain of 0.10 followed by Maximum temperature (Fig 1).

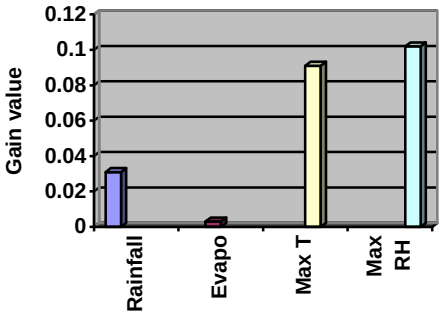
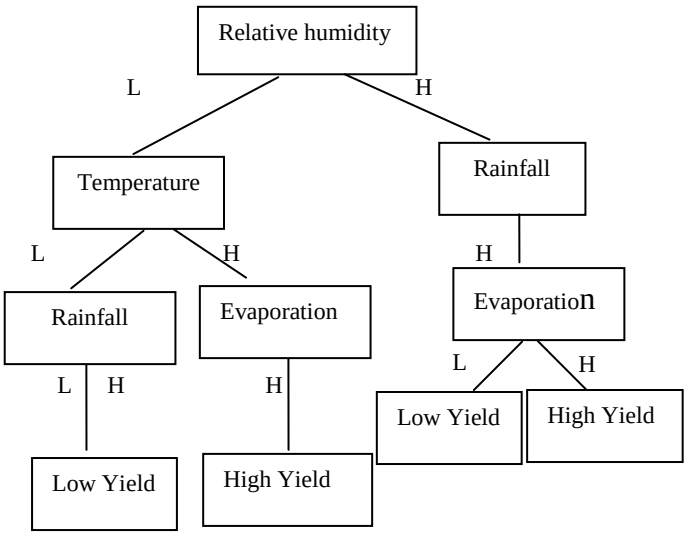


Fig 1. Gain obtained for soybean yield influencing parameters

A decision tree was constructed for soybean crop taking into consideration of Relative humidity as a major influencing parameter on the crop yield.

Decision tree for influence of climatic factors on soybean yield



Rule 1: RH = L, Temperature = L, Rainfall = L , Yield = L
Rule 2: RH = L, Temperaturel = H, Evaporation = H , Yield = H
Rule 3: RH = L, Temperature=L, Rainfall = H, Yield = Low
Rule 4: RH = H, Rainfall = H, Evaporation= L, Yield = L
Rule 5: RH = H, Rainfall = H, Evaporation= H, Yield = H

4. CONCLUSION

The present study is based on decision tree to assess the influence of climatic factors on the soybean crop yield. The study demonstrated the potential use of this process for extracting useful information from existing secondary data of climatic factors. The decision trees suggested there exists a correlation between climatic factors and soybean crop productivity and these variables influence on the soybean crop productivity were confirmed from the rule accuracy and Bayesian classification. The salient conclusions of the present study are:

i) The decision tree analysis indicated that the productivity of soybean crop was mostly influenced by Relative humidity followed by temperature and rainfall.

ii) The decision tree from the present study are fast to execute and much to be desired as representations of knowledge interpretations.

iii) The rules formed from the decision tree are helpful in predicting the conditions responsible for the high or low soybean crop productivity under given climatic parameters.

5. ACKNOWLEDGEMENTS

The first author would like to extend her heartfelt gratitude to Vice Chancellor, MGCGV, Chitrakoot for giving admission to pursue doctoral programme from the university. Thanks are also due to Director, CIAE for extending the facilities to carryout the research activities in the Institute. All the help received from the staff of the University and the Institute is duly acknowledged.

6. REFERENCES

[1] Abdullah, A., Brobst, S and M.Umer M. 2004. The case for an agri data ware house: Enabling analytical exploration of integrated agricultural data. Proceedings of IASTED International Conference on Databases and Applications. Austria. Feb.

[2] Abdullah, A., Brobst, S, Pervaiz.I., Umer M.,and A.Nisar. 2004. Learning dynamics of pesticide abuse through data mining. Proceedings of Australian Workshop on Data Mining and Web Intelligence, New Zealand, January.

[3] Abdullah, A., Bulbul.R., and Tahir Mehmood. 2005. Mapping nominal values to numbers by data mining spectral properties of leaves. Proceedings of 3rd International Symposium on Intelligent Information Technology in Agriculture. Beijing, China. Oct, 2005.

[4] Basak J., Sudharshan, A., Trivedi D. and M.S.Santhanam . 2004. Weather Data Mining Using Independent Component Analysis. J. of Machine Learning Research 5: 239-253.

[5] Chi-Chung LAU and Kuo-Hsin HSIAO, 2005. Bayesian Classification For Rice Paddy interpretation. Paper presented in Conference on data mining held at China Tapei. December, 2005.

[6] Cunningham S.J., and G. Holmes. 2005. Developing innovative applications in agriculture using data mining. Proceedings of 3rd International Symposium

on Intelligent Information Technology in Agriculture. Beijing, China. Oct, 2005.

[7] Han J and M Kamber, 2009. Data Mining Concepts and Techniques, Second Edition. Elsevier Publication. 285-305.

[8] Kiran Mai, C., Murali Krishna, I.V., and A.Venugopal Reddy, 2006. Data Mining Of Geo-spatial Database For Agriculture Related Application. Proceedings of Map India. New Delhi.

[9] Nain A. S., Dadhwal,V. K. and T. P. Singh, 2002 Real time wheat yield assessment using technology trend and crop simulation model with minimal data set. Current Science. 82(10): 1255-1258.

[10] Quinlan, J.R. (1979). Discovering rules by induction from large collections of examples. In D. Michie (Ed.), *Expert systems in the micro electronic age.* Edinburgh University Press.

[11] Quinlan, J.R. (1983a). Learning efficient classification procedures and their application to chess endgames. In R.S. Michalski, J.G. Carbonell & T.M. Mitchell, (Eds.), *Machine learning: An artificial intelligence approach.* Palo Alto: Tioga Publishing Company.

[12] Rish, I., Hellerstein, J., and T. Jayram. An analysis of data characteristics that affect naive Bayes performance. Technical Report RC21993, IBM T.J. Watson Research Center, 2001.

[13] S. Diplaris, S., Symeonidis' A.L., Mitkas, P.A., Bano, G. and Z. Abas, 2001. A decision-tree-based alarming system for the validation of national genetic evaluations. Computers in agriculture. 52: 21-35.

[14] Sanghi A, R Mendelsohn, and A Dinar. The climate sensitivity of Indian agriculture, In Measuring the Impact of Climate Change on Indian Agriculture, edited by A. Dinar, et al. Washington DC: World Bank. [World Bank Technical Paper No. 402] 1998.

[15] Singh Gyanendra and Hukum Chandra, 2002. Production and economic factors growth in Indian Agriculture. Technical Bulletin no. CIAE/2002/91: 1-216.