2014 International Conference on Computer Communication and Informatics (ICCCI -2014), Jan. 03 – 05, 2014, Coimbatore, INDIA

Machine learning approach for forecasting crop yield based on climatic parameters

S.Veenadhari
Ph.D.Scholar
MGCGV, Chitrakoot
Madhya Pradesh
veenadhari1@gmail.com

Dr. Bharat Misra
Associate Professor
MGCGV, Chitrakoot
Madhya Pradesh

Dr. CD Singh
Senior Scientist
CIAE, Bhopal
Madhya Pradesh

Abstract—With the impact of climate change in India, majority of the agricultural crops are being badly affected interms of their performance over a period of last two decades. Predicting the crop yield well ahead of its harvest would help the policy makers and farmers for taking appropriate measures for marketing and storage. Such predictions will also help the associated industries for planning the logistics of their business. Several methods of predicting and modeling crop yields have been developed in the past with varying rateof success, as these don't take into account characteristicsoftheweather, and aremostly empirical. In the present study a software tool named 'Crop Advisor' has been developed as an user friendly web page for predicting the influence of climatic parameters on the crop yields.C4.5 algorithm is used to find out the most influencing climatic parameter on the crop yields of selected crops in selected districts of Madhya Pradesh. This software provides an indication of relative influence of different climatic parameters on the crop yield, other agro-input parameters responsible for crop yield are not considered in this tool, since, application of these input parameters varies with individual fields in space and time.

Key words: Climate, agricultural productivity, C4.5 alogarithm, prediction

I.INTRODUCTION

Crop production is a complex phenomenon that is influenced by agro-climatic input parameters. Agriculture input parameters varies from field to field and farmer to farmer. Collecting such information on a larger area is a daunting task. However, the climatic information collected in India at every 1sq.m area in different parts of the district are tabulatedby Indian Meteorological Department. The huge such data sets can be used for predicting their influence on major crops of that particular district or place. There are different forecasting methodologies developed and evaluated by the researchers all over the world in the field of agriculture or associated sciences. Some of the such studies are : Agricultural researchers in Pakistan have shown that attempts of crop yield maximization through pro-pesticide state policies have led to a dangerously high pesticide usage. These studies have reported negative correlation between pesticide usage and crop yield[1]. In their study they have shown that how data mining integrated agricultural data including pest scouting, pesticide usage and meteorological data are useful for optimization of pesticide usage. Thematic information related to agriculture which has spatial attributes was reported in one of the study[6]. Their study aimed at discerning trends in agriculture production with references to the availability of inputs. K- means method was used to perform forecasts of the pollution in the atmosphere [4], the k nearest neighbor was applied for simulating daily precipitations and other weather variables [11], and different possible changes of the weather scenarios are analyzed using SVMs[13]. Data mining techniques are often used to study soil characteristics. As an example, the k-means approach is used for classifying soils in combination with GPS-based technologies [14]. Apples were checked using different approaches before sending them to the market.[9], uses a k-means approach to analyze color images of fruits as they run on conveyor belts. [12] uses X-ray images of apples to monitor the presence of water cores, and a neural network is trained for discriminating between good and bad apples. Spatial data mining introduced especially decision tree algorithm applying to agriculture land grading[15]. He combined spatial data mining techniques with expert system techniques and applied them to establish an intelligent agriculture landgrading information system. The author adopted decision tree C4.5 algorithm and implement with Mo2.0 and VC++6.0 to build agriculture land grading expert system. The study showed the advantages of this method in addressing problems in land grading. A decision tree classifier for agriculture data was proposed [5].This new classifier uses new data expression and can deal with both complete data and in complete data. In the experiment,10-fold cross validation

978-1-4799-2352-6/14/$31.00 ©2014 IEEE

method is used to test dataset, horse-colic dataset and soybean dataset. Their results showed the proposed decision tree is capable of classifying all kinds of agriculture data. Data mining technique for evolution of association rules for droughts and floods in India was applied using climate inputs[2]. In their study, a data-mining algorithm using the concepts of minimal occurrences with constraints and time lags was used to discover association rules between extreme rainfall events and climatic indices. Rainfall events were forecasted the using data mining techniques[7].The occurrence of prolonged dry period or heavy rain at the critical stages of the crop growth and development may lead to significant reduction in crop yield. Sugarcane yield was estimated in Brazil, using 10-day periods of SPOT vegetation NDVI images and meteorological data [3]. Data Mining approach based on Spatio-Temporal data to forecast irrigation water demand[8]. A set were prepared containing attributes obtained from meteorological data, remote sensing images and water delivery statements. In order to make the prepared data sets useful for demand forecasting and pattern extraction data sets were processed using a novel approach based on a combination of irrigation and data mining knowledge. Decision tree techniques were applied to forecast future water requirement.

II. METHODOLOGY

The present study was aimed to develop a web site for finding out the influence of climatic parameters on crop production in selected districts of Madhya Pradesh. The selection of districts has been made based on the area under that particular crop. Based on this criteria first top five districts in which the selected crop area is maximum were selected. The crops selected in the study is based on the predominate crops in the selected district. The selected crop includes: Soybean, Maize, Paddy and Wheat. The yield of these crops was tabulated for continuous 20 years by collecting the information from secondary sources. Similarly for the corresponding years climatic parameters such as Rainfall, Maximum & Minimum temperature, Potential Evapotranspiration, Cloud cover, Wetday frequency were also collected from the secondary sources. The methodology adopted for analysis includes for values above the threshold were considered as one child and the remaining as another child. It also handles missing attribute values. In pseudo code, the general algorithm for building decision trees is:

- Check for base cases
- For each attribute *a* : find the normalized information gain from splitting on *a*
- Let *a_best* be the attribute with the highest normalized information gain
- Create a decision *node* that splits on *a_best*
- Recurse on the sublists obtained by splitting on *a_best*, and add those nodes as children of *node.*

Let S be a set of training samples, where the class label of each sample is known. Each sample is in fact a tuple. One attribute is used to determine the class of training samples. Suppose that there are m classes. Let S contain s_i samples of class C_i for I= 1.., m. An arbitrary samples belongs to class Ci with probability si/s, where s is the total number of samples in set S. The expected information needed to classify a given sample is

$$I(s_1, s_2, \ldots, s_m) = -\sum_{i=1}^{m} \frac{s_i}{s} \log_2 \frac{s_i}{s}$$

An attribute A with values $\{a_1, a_2, \ldots, a_v\}$ can be used to partition S into the subsets $\{S_1, S_2, \ldots, S_v\}$, where S_j contains those samples in S that have value a_j of A. Let $S_j$contain s_{ij} samples of class Ci. The expected information based on this partitioning by A is known as the entropy of A. It is the weighted average:

$$E(A) = \sum_{j=1}^{v} \frac{s_{1j} + .., s_{mj}}{s} I(s_{1j}, \ldots s_{mj})$$

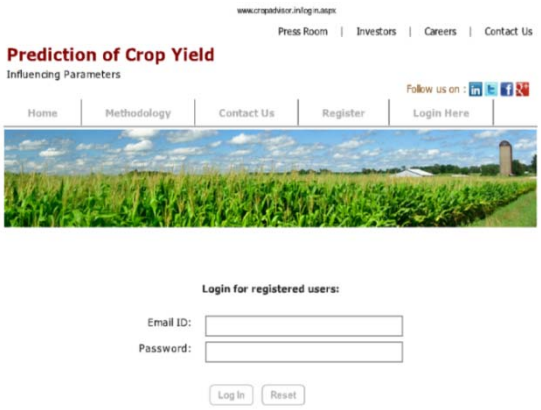
The information gain obtained by this portioning on A is defined by

$$Gain(A) = I(s_1, s_2, \ldots, s_m) - E(A)$$

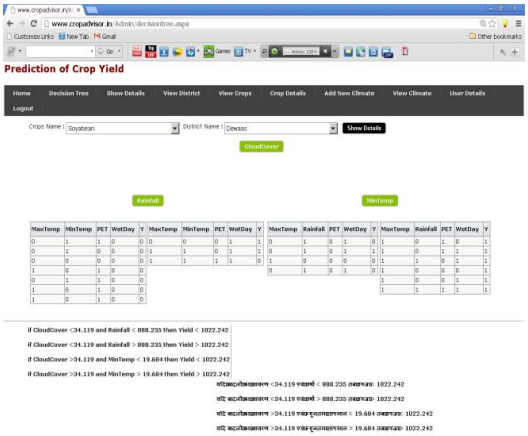
In this approach to relevance analysis, we can compute the information gain for each of the attributes defining the samples in S. The attribute with the highest information gain is considered the most discriminating attribute of the given set. By computing the information gain for each attribute, we therefore obtain a ranking of the attributes. This ranking can be used for relevance analysis to select the attributes to be used in concept description.

III. RESULTS AND DISCUSSION

A web based software has been developed in C# language in .net platform. The backend used is sql server 2008. On the home page of the web site (www.cropadvisor.in) the methodology adopted in the study, contact information of the administrator, new user registration and registered users login are appeared. For the registered users the window is displayed as



This website is designed as an interactive software tool for predicting the influence of climatic parameters on the crop yields. C 4.5 algorithm is used to find out the most influencing climatic parameter on the crop yields of selected crops in selected districts of Madhya Pradesh. This software provides an indication of relative influence of different climate parameters on the crop yield, other agro-input parameters responsible for crop yield are not considered in this tool, since, and application of these input parameters varies with individual fields in space and time. Based on the C 4.5 alogarithm, decision tree and decision rules have been developed, which are displayed when icon decision tree is selected. The screen shot of the same appear as:



Using the developed software the influence of climatic parameters on crop productivity in selected districts of Madhya Pradesh was carried out for predominant crops. For Soybean crop in all the selected districts, the most influencing parameter was found to be cloud cover, for paddy crop it was found as rainfall, for maize crop it was maximum temperature and for wheat crop the minimum temperature.

In the present study only the climatic parameters were considered in predicting the crop yield, though, the crop yield is influenced by many other input parameters such as irrigation, fertilizer application, pesticide application etc. This is due to paucity of such information on district wise resulted in developing a model, which can approximately predict the crop yields knowing the climatic parameters, as this will facilitate the policy makers to decide on the buffer stock of the grains, fixing minimum support price etc. Therefore, the decision rules that were framed from the developed software was used for validation of the software by predicting the yields of selected crops in all the selected district with the observed values. The prediction accuracy was also worked out by comparing the predicted yield with the observed yields. For each crop the validation of the developed software has been carried out, however for one crop (soybean) the decision rules and validation accuracy of results were presented in the present paper.

The decision rules developed based on the model for soybean crop in Dewas district are:

i) if Cloud Cover < 34.11days & Rainfall < 888.23 mm then Yield < 1022.24 kg/ha
ii) if Cloud Cover <34.11 days & Rainfall < 888.23 mm then Yield >1022.24 kg/ha
iii) if Cloud Cover >34.11 days & Minimum Temperature < 19.68^{o}C then Yield < 1022.24 kg/ha
iv) if Cloud Cover >34.11 days & Minimum Temperature > 19.68 ^{o}C then Yield > 1022.24 kg/ha

Based on the above decision rules the observed values of the most influencing parameters of this district are presented in Table 1.0.

Table 1.0 Prediction accuracy of developed model for soybean crop in Dewas district

Cloud cover, days	Rainfall, mm	Min. temp, ^{o}C	Observed Yield, kg/ha	Predicted Yield, kg/ha	Is prediction accurate
34.96	833.19	19.72	1092.7	>1022.24	Yes
33.96	1161.41	19.25	1060.8	>1022.24	Yes
34.04	938.17	19.83	1060.8	>1022.24	Yes
32.6	1033.26	20.06	1007.8	>1022.24	No
34.86	1067.67	19.27	1022.2	<1022.2	Yes
34.7	833.15	20.24	1247.1	>1022.24	Yes
35.17	962.4	19.77	1146.6	>1022.24	Yes
31.76	693.67	19.93	1009.8	<1022.24	Yes
35.23	694.35	19.7	1160.1	>1022.24	Yes
33.14	692.42	20.08	985.8	<1022.24	Yes
34.33	1168	18.98	1099	<1022.24	No
33.81	824	19.24	925	<1022.24	Yes
34.13	878	19.76	1275	>1022.24	Yes
33.86	851	19.49	912	<1022.24	Yes
33.85	802	20.1	907	<1022.24	Yes
34.49	915	20.14	1023	>1022.24	Yes
34.53	687	19.5	880	<1022.24	Yes
34.98	1293	19.31	780	<1022.24	Yes
34.02	709	19.65	920	<1022.24	Yes
33.96	728	19.66	930	<1022.24	Yes

Out of 20 years of data the predictions were correct in 18 years and were incorrect in two years indicating the prediction accuracy of the developed model at 90 per cent in case of soybean in Dewas district. Similar analysis were carried out for all the selected crops and districts, and based on the results obtained the overall accuracy of the developed model are presented in Table 2.0.

The web based software developed for predicting the crop yield from the given input of climatological parameters indicated a clear trend of each crop being predominantly influenced by a particular climatic parameter. The average of accuracy obtained under a particular crop in different districts were averaged and the prediction accuracy of developed model for different crops are presented in table 2.0.

Table 2.0. Prediction accuracy of developed model for different crops.

S.No.	Name of the Crop	Average prediction accuracy, %
1	Soyabean	87
2	Paddy	85
3	Maize	76
4	Wheat	80

The prediction accuracy of the developed model varied from 76 to 90 per cent for the selected crops and selected districts. Based on these observations the overall prediction accuracy of the developed model is 82.00 per cent. With a high prediction accuracy the developed model can be used by the policy makers in arriving at a policy decision well in advance i.e., before the harvest of the crop.

IV. CONCLUSIONS

The present study demonstrated the potential use of data mining techniques in predicting the crop yield based on the climatic input parameters. The developed webpage is user friendly and the accuracy of predictions are above 75 per cent in all the crops and districts selected in the study indicating higher accuracy of prediction. The user friendly web page developed for predicting crop yield can be used by any user their choice of crop by providing climatic data of that place.

ACKNOWLEDGEMENTS

The first author would like to extend her heart felt gratitude to Vice Chancellor, MGCGV, Chitrakoot for giving admission to pursue Doctoral program from the university. Thanks are also due to Director, CIAE, Bhopal 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.

REFERENCES

[1]. 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.

[2]. Dhanya,C.T. and D. Nagesh Kumar, 2009. Data mining for evolution of association rules for droughts and floods in India using climate inputs. J. of Geo. Phy.Res.114:1-15.

[3]. Fernandes F.L., Jansle V.R., Rubens Augusto and Camargo L.(2011). Sugarcane yield estimates using time series analysis of spot vegetation images. Sci. Agric. (Piracicaba, Braz.)vol.68no.2

[4]. Jorquera H, Perez R, Cipriano A, Acuna G(2001). Short term forecasting of air pollution episodes. In. Zannetti P (eds) Environmental Modeling 4. WITPress, UK.

[5]. Jun Wu, Anastasiya Olesnikova, Chi- Hwa Song, Won Don Lee (2009).The Development and Application of Decision Tree for Agriculture Data. IITSI :16-20.

[6]. Kiran Mai,C., Murali Krishna, I.V., and A.VenugopalReddy,2006.Data Mining o f Geospatial Database for Agriculture Related Application. Proceedings of Map India. New Delhi. ((http://www.gisdevelopment.net/proceedings/mapindia/2006/agriculture/mi06agri_124.htm).

[7]. Kannan, M. Prabhakaran S and P. Ramachandran (2011).Rainfall forecasting using data mining technique. International Journal of Engineering and Technology Vol.2 (6), 2010, 397-401.

[8]. Khan Mohammad A.,Md. Zahid-ul Islam and Mohsin Hafeez (2011). Evaluating the Performance of Several Data Mining Methods for Predicting Irrigation Water requirement. Proceedings of the Tenth Australasian Data Mining conference(AusDM2012),Sydney, Australia.199-207.

[9]. Leemans,V.,Destain,M.F.,2004.A real-time grading method of apples based on features extracted from defects. J. Food Eng.61, 83–89.

[10]. Quinlan,J.R.(1985b). Decision trees and multi-valued attributes. In J.E. Hayes & D.Michie (Eds.),*Machine intelligence 11*. Oxford University Press (in press).

[11]. Rajagopalan B. Lall U (1999) A k- nearest-neighbor daily precipitation and other weather variables. WatResResearch35(10) :3089 – 3101.

[12]. Shahin,M.A.,Tollner,E.W.,McClendon,R.W. Arabnia, H.R.,(2002). Apple classification based on surface bruises using image processing and neural networks. Trans. ASAE 45, 1619–1627.

[13]. Tripathi S, Srinivas VV, Nanjudiah RS (2006).Down scaling of precipitation for climate change scenarios: a support vector machine approach, J. Hydrology 330-337.

[14]. Verheyen, K., Adrianens, M. Hermy and S.Deckers(2001). High resolution continuous soil classification using morphological soil profile descriptions. Geoderma, 101:31-48.

[15]. Zelu Zia (2009). An Expert System Based on Spatial Data Mining used Decision Tree for Agriculture Land Grading. Second International Conference on Intelligent Computation Technology and Automation.Oct10-11,China