Review: Role of Data Mining in Agriculture Yield Analysis

Rupinder Singh[*] and Gurpreet Singh[**]
[*]Research Scholar, Department of Computer Engineering, Punjabi University, Patiala, (India)
rupi85aujla@gmail.com
[**]Assistant Professor, Department of Computer Engineering, Punjabi University, Patiala, (India)
Gurpreet.1887@gmail.com

Abstract: Use of data mining has been increased in agriculture field due to enhancement in the technology. Agriculture yield analysis is a very complex and vast research area as it deals with large data sets including different factors viz. yields of various crops, meteorological parameters affecting crop yields, diseases, pests etc. Data mining methods extract meaningful patterns from those large data sets. This paper presents a review on how different data mining techniques are helpful to study the impact of different meteorological parameters on various crop yields.

Keywords: Data Mining, Agriculture yield estimation, Regression, Decision Tree Induction.

Introduction

Agriculture plays very effective and important role in the Indian economic system. Almost 43% of geographical area in India has been covered by agriculture sector. Agriculture is very important occupation for us as it provides not only the food but also the raw materials to other industries. More than 75% of population in India depends on agriculture and agriculture related tasks for their livelihood [15]. In 2015 – 16, agriculture's share in Gross Domestic Product (GDP) of India was 17.4 % as compared to 18.3 % in 2013 – 14 [9]. Still agriculture is the largest contributor to India's GDP even after decrease in share of agriculture sector. The prediction of agriculture yield is one of important steps taken by producers and policy makers to monitor the growth of agriculture sector.

Agriculture yield may depend on many factors that are independent of one another includes geographical, climatic, biological factors and economic policies. The proper management of these factors is essential to get significant results for agriculture output. The crop yield also depends on other factors like pests, diseases, weeds, time of harvesting etc [1]. Before the crop growing season, the main motive of a farmer is to know about how much yield could be expected under available circumstances. Agriculture research oriented areas contain huge amounts of data. The main task of data mining is to uncover the meaningful patterns from such vast amounts of data. It aims at discovering such knowledge from data that should be both useful and important for farmers.

Data mining discovers all those patterns which should be useful, valid, novel and easily understood by humans [10]. For discovering such patterns, various data mining techniques are used. Data mining techniques widely used in various areas viz. fraud detection, market research, production control, medical diagnosis, meteorological analysis, agriculture, customer retention (Holding) and science exploration [10]. Data mining techniques mainly divided into two types of tasks namely predictive and descriptive.

Predictive Tasks

Predictive data mining techniques are supervised learning techniques. These techniques are used to generate models from class labeled data. Such produced models can be used for classification or prediction. It includes data mining techniques like classification, regression, time series analysis etc for data analysis. Classification is a process of organizing data into predefined categories with class labels where as regression tries to map a data item to a real valued prediction variable. Time series analysis technique includes prediction of future values based on the discovery of similar patterns over a particular time period.

Descriptive Tasks

To derive patterns that summarizes the underlying relationship between data. It is used to find human-interpretable patterns describing the data. These techniques can be used to generate useful patterns from unlabeled data. Such techniques include clustering, association rules, sequential pattern discovery for data analysis. Clustering is a technique of dividing data into different meaningful subsets called clusters. In clustering method, there are no such predefined classes occurred like in classification. Association rule mining method is one of the useful techniques of data mining to discover interesting and meaningful patterns that frequently occurs together in data. Sequential pattern discovery method includes the extraction of frequently occurring patterns in the data. It compares the different sequences and recovers the missing sequence numbers.

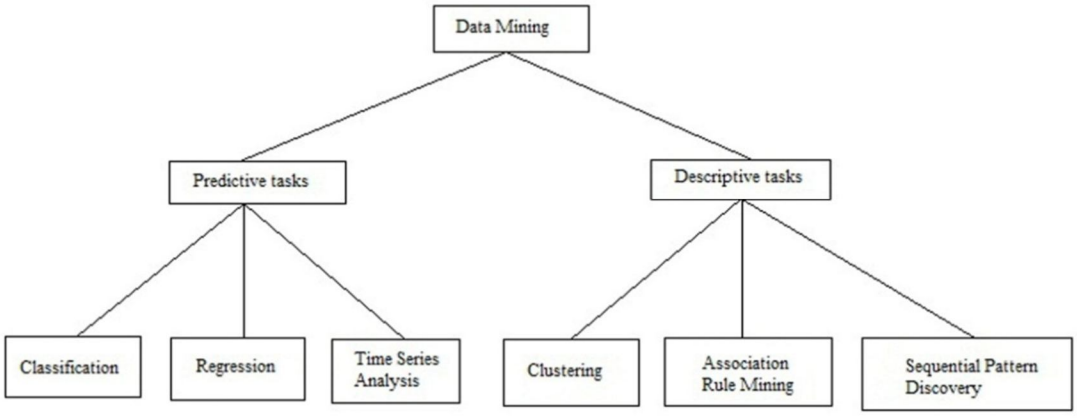

Fig. 1.1: Classification of data mining tasks

Knowledge Discovery from Data (KDD)

Some people use data mining as a synonym for Knowledge Discovery from Data, or KDD. On the other hand, others view data mining as a step in the process of knowledge discovery [10]. Knowledge discovery process consists of an iterative sequence of the following steps:

1. *Data cleaning*: It is the first step of KDD during which noise and inconsistent data is removed from given data sets.
2. *Data integration*: During data integration, multiple data sources are combined for further analysis.
3. *Data selection*: In this step, data relevant to the analysis task are retrieved from the databases and other information repositories.
4. *Data transformation*: During data transformation, data sets are transformed into forms appropriate for mining by performing various operations on them.
5. *Data mining*: Data mining is a process where different methods and algorithms are applied in order to discover some useful data patterns.
6. *Pattern evaluation*: During this step, some interesting patterns are identified by using different interestingness measures.
7. *Knowledge presentation*: It is the last step of KDD during which visualization and information representation techniques are applied in order to present the mined data to the user.

Steps 1 to 4 constitute the process of data preprocessing, where the data is prepared for mining process.

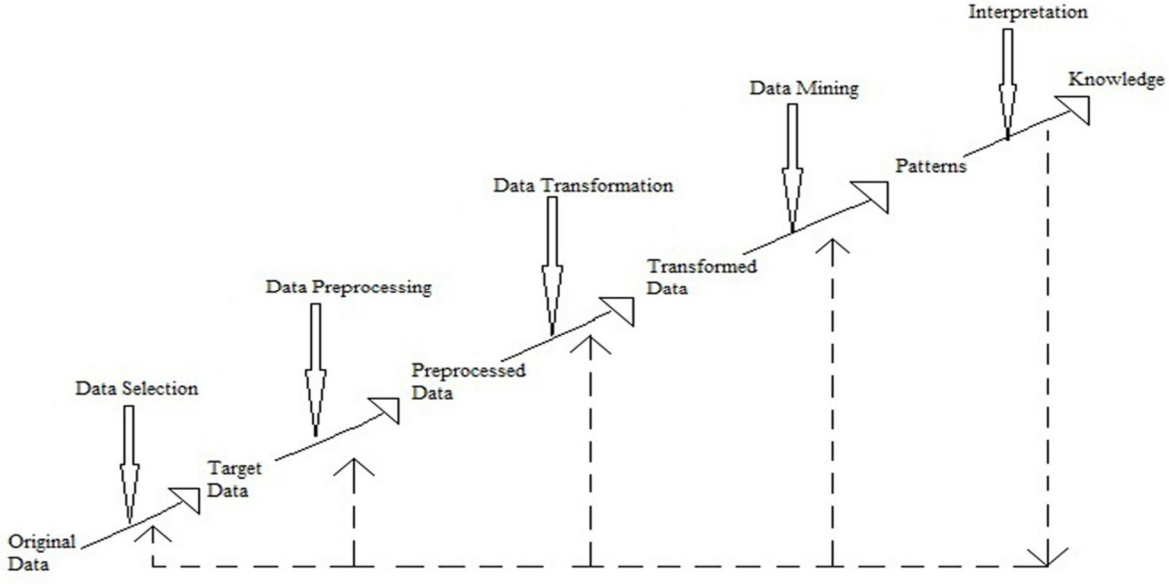

Fig. 1.2: Knowledge discovery of data

Applications of data mining techniques in agriculture

There are different data mining techniques and algorithms are available for crop yield analysis and estimation. An effective methodology for crop yield prediction can be built up by using following data mining methods.

Regression

Regression allows us to model the relationship between two or more variables using simple mathematical techniques. Regression method works on two types of variables viz. independent variables and dependent variables. In practical life, regression analysis are applied to predict profit, sales, credit rates, house values, crop yield, temperature, distance between two or more points etc. A regression model that predicts the crop yield could be developed based on observed data for many yields of that particular crop over a period of time. In addition to the value, the data might track the sowing area, temperature, rainfall, humidity, fertilizers used, number of pesticides and spray applied and so on [13].

Association Rule Mining

Association rule mining method is one of the most important and useful method of data mining to discover interesting and meaningful patterns among large amount of data. Association rules are in the form of IF – THEN statements which help to find the desired relationships occurred between the various instances of data stored in data warehouses and other information repositories. These rules are applied in various areas viz. medical diagnosis, logistics, marketing and agriculture. In agriculture research domain, association rule mining helps to discover useful information and generates important rules about crop yields based on the relationships between different crops and soil parameters [8]. For this purpose, the various association rule mining algorithms like Apriori, Predictive Apriori and FP Growth algorithms are applied for agriculture yield data analysis.

K Nearest Neighbors (KNN)

KNN is a classification and regression method mostly used for pattern recognition and statistical estimates. KNN classifies the objects based on the distance functions. In case of continues variables Euclidean, Manhattan and Minkowski distance functions are used for calculation whereas in case of categorical variables or binary data, hamming distance is used for statistical calculation. During classification, KNN algorithm provides a class having highest frequency count among K most similar instances as an output. But in case of regression, KNN provides the output based on the mean or median of K most similar instances. KNN applied for resampling of weather variables in order to design a K Nearest Neighbors simulator for daily precipitation and other climate parameters [2].

K Means Clustering

K Means Clustering is an unsupervised learning algorithm of clustering technique. It partitions given instances of data into K clusters by using mean of cluster as key parameter. Each instance present in cluster is nearest to the mean of that cluster. Major application areas of K Means Clustering include image processing, market research, pattern recognition, medical data analysis, meteorological data analysis and agriculture. In agriculture research field, K Means Clustering method is capable to partition the samples of crop yields and weather parameters into different clusters which are helpful for agriculture yield analysis [4].

Decision Tree Induction

Decision tree method includes the learning of decision trees from class-labeled training data sets. A decision tree is a flowchart-like tree structure, where each internal node represents a test on an attribute, each branch denotes an outcome of the test, and each leaf node represents a class label. The uppermost node of the tree represents the root node. The attribute values of a given data sample are tested against decision tree for classifying that unknown data sample. Decision tree induction method has been used in the field of biomedical engineering, financial analysis, manufacturing and production. This technique can be applied on agricultural data set to predict the impact of climate parameters on crop productivity based on the relationship between crop and weather parameters [17].

Support Vector Machine (SVM)

SVM is a supervised learning method for classification of both linear and non linear data. It uses a non linear mapping to transform original training data into a higher dimension [10]. SVM classifies the data by finding the hyperplane that maximizes margin width between any two classes. SVM technique has been used in many fields includes bioinformatics, multimedia, artificial intelligence, pattern recognition, agriculture and so on. A SVM based downscaling model applied to obtain the future projections of precipitations for meteorological sub divisions in India [16].

Literature survey

D. B. Lobell and C. B. Field [3] discussed about the impacts of recent warming on the production of major crops in the world. Average global yields of six major crops for time period of 42 years were taken into consideration for results

generation. Multiple linear regressions have been performed using global yields as response variables and climate parameters viz. minimum & maximum temperature and rainfall as predictor variables. Analyses suggest that increased atmospheric temperature had negative impact on global yields of several major crops. There was a clear decline in yields of wheat, maize and barley crops with respect to temperature rise in the past. As climate changes, farmers would accommodate such cropping systems in order to minimize the negative impacts of warming.

D. R. Mehta, A. D. Kalola, D. A Saradava and A. S. Yusufzai [5] have analyzed the impacts of rainfall variability on various crop yields. The weekly rainfall data for 39 years and district average yield data for 36 years has been used for results generation. A yield estimation model has been developed with the help of correlation and regression methods by using rainfall as an independent variable and crop yield as dependent variable. Positive correlation was found between rainfall and yields of groundnut, pearl millet and sorghum.

D. W. Parvin, S. W. Martin, F. Cooke, Jr., and B. B. Freeland, Jr. [6] studied the effect of harvest season rainfall on cotton yield at 22 locations in the Delta area of Mississippi from year 1991 to 1993 and 2002. For this purpose Regression analysis were performed to estimate the relationship between yield as the dependent variable and time and rainfall as independent variable. After the analysis, results indicate that during the harvest season increase in rainfall decreases the cotton yield and rainfall during the late season results in greater yield reduction.

E. M. Adamgbe and F. Ujoh [7] discussed about the impact of rainfall on maize yield in Gboko, Nigeria. The data collected over 30 years was analyzed using mean, correlation and regression methods to portray the relationship between rainfall and maize yield. Results indicate that delay in the rainy season caused shortening of crop growing season thus reducing maize yield.

F. Khan and D. Singh [8] presented the implementation of association rule mining methodology for analysis of agricultural data set in order to generate rules to discover the relationships between different crop yields. The data sets of five different crops from Bhopal district in Madhya Pradesh were collected for analysis work. The parameters like soil type, PH value of the soil and cropping season were into consideration for results generation. The results generated through Apriori algorithm are further compared with results obtained by FP Growth method.

P. Gwimbi and T. Mundoga [11] have been presented the impact of climate change on cotton production under rain fed conditions in Gokwe, Zimbabwe. The dataset was taken of 25 years and a survey of 50 farmers in Gokwe district for proposed work. Significant Climate pattern were generated using rainfall and temperature data were statistically correlated to cotton yield using Statistical Package for the Social Sciences (SPSS) software package. This correlation provided evidence of the relationship between rainfall and temperature variability and cotton production over that time period. The results generated by SPSS tool indicate that rainfall has positive impact on cotton yield but increase in temperature follows the considerable decrease in the cotton yield.

R. Dehgahi, A. Joniyas and M. D. Latip [12] presented in their paper about rainfall distribution and temperature effects on wheat yield in Torbat Heydrei during 1993 to 2008. They stated that wheat yield is a result of the height of plants, the number of productive tillers, the number of grains per spike etc. The data sets were analyzed using Excel and Minitab (Anova) in order to determine the effect of rainfall and temperature on wheat yield. The results of variance analysis between rainfall and wheat yield showed a strong relationship between rainfall and wheat yield as high temperature and low rainfall causes reduction in wheat yield.

S. Kaul [13] suggested that crop yield depends on both climate variables and social factors. The results obtained from data showed that climatic variables responsible for 19% of the yield change while the social variables responsible for 74% of the yield change. Regression techniques applied to study the impact of climate change on the productivity of rice and jowar crops. Results showed that high rainfall and extreme variations in temperature would have negative effect on yield of these crops thus reducing farmers' income. To discover the real effects of climate and social factors on productivity of jowar crop, further analyses are required.

S. S. Hussain, M. Mudasser, M. M. Sheikh and N. Manzoor [14] discussed about the climate changes in mountain areas of Pakistan and their consequences on fresh water resources and agriculture yield. Regression techniques applied to climate data collected over the period of 30 years. Results obtained after analysis indicate an increase in the temperature for both winter and monsoon seasons over the years. This increase in temperature not only has positive impact on agriculture yield but also will affect fresh water resources of the country in long term.

S. Veenadhari, Dr. B. Mishra and Dr. C. D. Singh [17] presented in their paper about soybean productivity modeling using Decision Tree Algorithms. Bayesian classification and rule accuracy together with decision trees applied to study the impact of climate variables on crop productivity. Data sets of meteorological data of Bhopal district for 20 years were collected for results generation. Results obtained after analysis indicate that the productivity of soybean crop was mostly influenced by Relative humidity followed by rainfall and temperature variables.

Table 3.1: Various techniques and parameters used for agriculture yield prediction

S.No.	Authors	Crop under study	Parameters under study	Techniques applied
1	D. B. Lobell and C. B. Field [3]	Wheat, Rice, Maize, Soybean, Barley and Sorghum	Maximum and minimum temperature, rainfall	Multiple linear regression
2	D. R. Mehta, A. D. Kalola, D. A Saradava and A. S. Yusufzai [5]	Groundnut, Pearl millet, Sorghum and Cotton	Rainfall	Correlation and Regression
3	D. W. Parvin, S. W. Martin, F. Cooke, Jr., and B. B. Freeland, Jr. [6]	Cotton	Rainfall	Regression
4	E. M. Adamgbe and F. Ujoh [7]	Maize	Rainfall	Correlation and Regression
5	F. Khan and D. Singh [8]	Jower, Bajra Rice, Soybean and Wheat	Soil type, ph value of soil and Crop growing season	FP-Growth algorithm and Apriori algorithm
6	P. Gwimbi and T. Mundoga [11]	Cotton	Rainfall and temperature	Statistical Package for the Social Sciences
7	R. Dehgahi, A. Joniyas and M. D. Latip [12]	Wheat	Temperature and rainfall	Variance analysis
8	S. Kaul [13]	Rice and Jowar	Maximum and minimum temperature, rainfall, fertilizers usage and human labor	Regression
9	S. S. Hussain, M. Mudasser, M. M. Sheikh and N. Manzoor [14]	Wheat and Barley	Temperature and rainfall	Regression
10	S. Veenadhari, Dr. B. Mishra and Dr. C. D. Singh [17]	Soybean	Rainfall, temperature, evaporation and relative humidity	Decision Tree and Bayesian classification

Discussion

An inclusive summary of 15 articles is presented in this paper. The existing review paper by A. A. Raorane and R. V. Kulkarni [1] covers the overview of data mining whole agriculture sector whereas we put more emphasis on studying the effect of meteorological parameters on different crops. Different data mining techniques and algorithms which have been used for agriculture yield analysis are presented in this paper. In cited literature, data selection was carried out independently by researchers to generate results. We presented the qualitative overview of effect of various parameters on different crops along with the details of techniques applied in the form of table. In addition to this, we also tried to summarize the applications of different data mining techniques in weather forecasting.

Conclusion

In the view of this, there are certain climate parameters responsible for variable crop yields. Various data mining techniques are available for analysis of different weather parameters with respect to different crop yields. By using these techniques one can build up a methodology for pre harvest crop forecasting. To find out the effect of meteorological parameters on a crop, a combination of two or more data mining algorithms can be applied to get better results.

References

[1] A. A. Raorane and R. V. Kulkarni, "Review- Role of Data Mining in Agriculture," *International Journal of Computer Science and Information Technologies*, ISSN: 0975-9646, Vol. 4 No. 2, pp. 270–272, 2013.

[2] B. Rajagopalan and U. Lall, "A K–Nearest-Neighbor Simulator for Daily Precipitation and Other Weather Variables," *Water Resources Research*, Vol. 35, No. 10, pp. 3089–3101, October 1999.

[3] D. B. Lobell and C. B. Field, "Global scale climate–crop yield relationships and the impacts of recent warming," *Environmental Research Letters*, Vol. 2, pp. 1-7, 2007.

[4] D. Ramesh and B. V. Vardhan, "Data Mining Techniques and Applications to Agricultural Yield Data," *International Journal of Advanced Research in Computer and Communication Engineering*, ISSN: 2319-5940, Vol. 2, Issue 9, pp. 3477-3480, 2013.

[5] D. R. Mehta, A. D. Kalola, D. A Saradava and A. S. Yusufzai, "Rainfall Variability Analysis and its Impact on Crop Productivity – A Case Study," *Indian Journal of Agricultural Research*, Vol. 36, No. 1, pp. 29-33, 2002.

[6] D. W. Parvin, S. W. Martin, F. Cooke, Jr., and B. B. Freeland, Jr., "Effect of Harvest Season Rainfall on Cotton Yield," *Journal of Cotton Science*, Vol. 9, pp. 115–120, 2005.
[7] E. M. Adamgbe and F. Ujoh, "Effect of Variability in Rainfall Characteristics on Maize Yield in Gboko, Nigeria," *Journal of Environmental Protection*, Vol. 4, pp. 881-887, 2013**.**
[8] F. Khan and D. Singh (2014), "Knowledge Discovery on Agricultural Dataset Using Association Rule Mining," *International Journal of Emerging Technology and Advanced Engineering*, Vol. 4 Issue 5, pp. 925-930, May 2014.
[9] India Economic Survey 2015-16 – Key Highlights. [Online]. Available: https://home.kpmg.com/content/dam/kpmg/pdf/2016/04/KPMG-Flash-News-India-Economic-Survey-2015-16%E2%80%93Key-Highlights-3.pdf
[10] J. Han and M. Kamber, *Data Mining: Concepts and Technologies*. San Francisco, CA: Morgan Kaufmann, 2006.
[11] P. Gwimbi and T. Mundoga, "Impact of Climate Change on Cotton Production under Rainfed Conditions: Case of Gokwe," *Journal of Sustainable Development in Africa*, ISSN: 1520-5509, vol. 12, No. 8, pp. 59-69, 2010.
[12] R. Dehgahi, A. Joniyas and M. D. Latip, "Rainfall Distribution and Temperature Effects on Wheat Yield in Torbate Heydarie," *International Journal of Science Research in Knowledge*, ISSN: 2322-4541, 2(Special Issue), pp. 121-126, 2014.
[13] S. Kaul. (2001). Bio-Economic Modelling of Climate Change on Crop Production in India. [Online]. Available: www.ecomod.org/files/papers/370.pdf
[14] S. S. Hussain, M. Mudasser, M. M. Sheikh and N. Manzoor, "Climate Change and Variability in Mountain Regions of Pakistan Implications for Water and Agriculture," *Pakistan Journal of Meteorology*, Vol. 2, Issue 4, pp. 75-90, November 2005.
[15] S. Thenmozhi and P. Thilagavathi, "Impact of Agriculture on Indian Economy," *International Research Journal of Agriculture and Rural Development*, ISSN: 2319-331X, Vol. 3, No. 1, pp. 96-105, December 2014.
[16] S. Tripathi et al., "Downscaling of Precipitation for Climate Change Scenarios: A Support Vector Machine Approach," *Journal of Hydrology*, 330, pp. 621-640, 2006.
[17] S. Veenadhari, Dr. B. Mishra and Dr. C. D. Singh, "Soybean Productivity Modelling using Decision Tree Algorithms," *International Journal of Computer Applications*, Vol. 27, No. 7, pp. 11-15, August 2011.