ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

IJARCCE
International Journal of Advanced Research in Computer and Communication Engineering
ICRITCSA
M S Ramaiah Institute of Technology, Bangalore
Vol. 5, Special Issue 2, October 2016

Comparative Study of Data Mining Techniques in Crop Yield Prediction

Perpetua Noronha[1], Divya .J[2], Shruthi .B.S[3]

Lecturer, Department of Computer Science, Mount Carmel College, Bengaluru, India[1]

Student, Department of Computer Science, Mount Carmel College, Bengaluru, India[2, 3]

Abstract: Agriculture is the field of interest in today's technology emerging world. It is the main occupation and backbone of our country. As India's population currently stands at 1.3 billion people and is projected to grow eight times of current population by 2024, its become a critical challenge for the farmers to feed the population. And also the various environmental changes in the developing world are posing an important threat to the agricultural economy. Hence food security enhancement requires the transition to agricultural production systems that are more productive. The need to incorporate Information technologies into the task of food production is very important. Crop yield prediction is one of the important factors that provide information for decision makers to maximize the crop productivity but it is a problem that needs to be solved based on available data. Data mining technology serves to be a better choice for this purpose and has become an interesting and recent research topic in agriculture to predict the crop yield. This paper presents a brief comparative study of various papers that deal with various techniques used to predict the crop yield. From the data that is readily available, the data mining techniques give a complete picture about the estimation of crop yield. Different data mining techniques that are in use for the crop yield estimation are K-Means, K-Nearest neighbor (KNN).

Keywords: K-Means, K-Nearest Neighbor (KNN), Support Vector Machine (SVM), Multiple Linear Regression (MLR).

I. INTRODUCTION

Data mining is a process of extracting important and useful information from large data sets. It is one of the important techniques used in crop yield prediction today. Agriculture in India is the center area for sustenance and dietary security. Global warming is projected to have significant impacts on conditions affecting agriculture which is a major drawback for the farmer to get back his profit. As a result, the rate of suicide among Indian farmers is increasing drastically causing devastating impact on the agricultural economy. In this regard the farmers would be interested to know the yield prior to the sowing of crops with many factors such as weather or temperature, rainfall, etc. This prediction could help the farmers in a wider range. Raw data, which is obtained from the history of crop yields is required for the prediction. These data play an important role in providing the supply chain operation for companies that require raw material that is used for agricultural produce. Agriculture depends on various independent factors such as climate, geography, political and economic factors. The yield of agriculture depends on climate, pesticides and harvest planning. Since the technology is emerging in all fields, the crop models and predictive tools could be element that is crucial to be expected. A producer not only grows crops but also large amount of data to which data mining techniques are applied and patterns that are of interest to the farmer are found.

For the researcher to get the values from the field he needs to conduct a survey and test the fertility of the soil and then define the samples. He should know the homogeneity of the soil to classify it into different strata. To know the homogeneity the soil fertility or fertility gradient at different parts of the field should be tested. Based on this homogeneity the soil is classified into different strata and the samples are selected randomly to analyze. The values that are derived would serve as the past history for which the classification and clustering techniques can be applied.

The two main categories of data mining are classification and clustering. To classify the unknown samples based on the information of classified samples, classification techniques are used. The classified samples used for classification are known as training set. The two important classification techniques which use these training sets to classify the unknown samples are Neural Networks and Support Vector Machines (SVM).

II. DATA MINING TECHNIQUES

A. Multiple linear regression

Multiple linear regression is method that is used to have a linear relationship between one or more independent variables and a dependent variable. The independent variable is used to estimate values for the dependent

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

IJARCCE

International Journal of Advanced Research in Computer and Communication Engineering

ICRITCSA

M S Ramaiah Institute of Technology, Bangalore

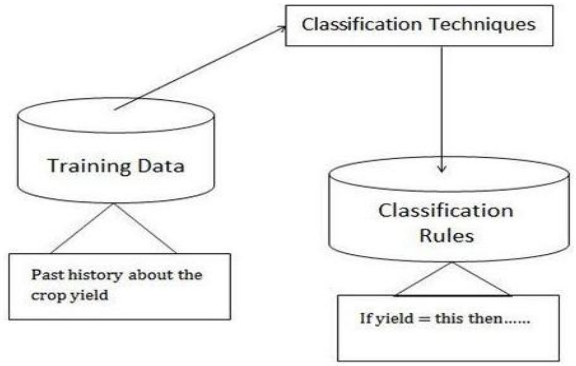
Vol. 5, Special Issue 2, October 2016

variables. MLR is a predictive analysis based on least squares and it is probably the most widely used method in climatology. The 3 major uses of MLR analysis is in casual regression, forecasting an effect, trend forecasting. Regression analysis focuses on the relationship between two variables whereas correlation analysis only focuses on the strength between the two or more variables.

B. K-Nearest Neighbor

K-Nearest Neighbor is a classification technique in which it is assumed that samples that are similar will have similar classification. The number of similar known samples used for assigning a classification to an unknown sample defines the parameter K.

In case if there is no history about the samples to be classified that is if the training set is not provided then clustering technique is used.



I. Diagram representing Classification Technique

C. K-Means Approach

The most important clustering technique is K-Means clustering. This technique is used to classify the data which have no previous knowledge about the data or the training set. The parameter K denotes the amount of clusters required to partition the data. [9]The idea of this clustering technique is, given K number of clusters we can define K centers, one for each cluster based on all samples belonging to a cluster. These centers must be placed far away from each other and then associate each sample to the cluster that has the closest centroid.

When no samples are left, the process of finding new K centers and assigning samples to the clusters that has the closest centroid is iteratively carried out until no longer the samples can change their clusters. In the research article [9], the researcher states that using this K-Means approach the Government could help the agricultural firms to increase one of their production practices such as acquiring new farmers by framing new profitable agriculture schemes based on crop yield prediction and eventually campaigning to different groups of farmers about a particular scheme based on the result of the grouping of various crops depending on common features.

D. Artificial neural network

Artificial neural network is one of the new data mining techniques that are based on biological neural processes of human brain. According to this technique once the neural network is trained it can predict the crop yield in similar patterns even if the past data include some errors. Even if the data is complex, multivariate, nonlinear this network gives the accurate results and also without any of underlying principles the relationship between them the output is extracted.

E. Support Vector Machines

Support Vector Machines (SVMs) are binary classifiers that will classify data samples in two disjoint classes. It is a technique in which two classes are linearly separable which is from a simplified case. [7]SVM can build a model that predicts whether a new example falls into category or the other.

A support vector machine is a concept in statistics and computer science for a set of related supervised learning methods that analyze data and recognize patterns used for classification and regression analysis. The SVM takes a set of input data and predicts for each given input which of two possible classes forms the input making the SVM a non probabilistic binary linear classifier.

F. Regression Model

Regression model are also used in crop yield prediction. Regression is mainly used for predicting about the future (not only crop yield). This model defines two variables independent and dependent variable. The value of the dependent variable can be predicted using the independent variable. Ex: In case of crop yield and soil, yield is dependent on the type of the soil. If that type of soil is suitable for that crop then the yield is high.

G. Biclustering Technique

Biclustering technique is one of the techniques used in data mining when the data is given in the form of rows and columns that is in the matrix form. Different types of bi-clustering algorithms are: Bicluster with constant values, Bicluster with constant rows, Bicluster with constant columns, Bicluster with coherent values.

Among these techniques K-Means, KNN techniques are suitable to predict the crop yield accurately. These above mentioned techniques require knowledge of statistics. Regression model is used when the researcher is having the clear picture about the type of variable whether it is dependent or independent variable. In case if there are more than two independent variables then multiple linear regression is preferred.
When the data is given in form of matrix the researcher is advised to use biclustering technique. If the data is based on biological neural processes of human brain then artificial neural network is advisable.

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

IJARCCE

International Journal of Advanced Research in Computer and Communication Engineering

ICRITCSA

M S Ramaiah Institute of Technology, Bangalore

Vol. 5, Special Issue 2, October 2016

III. LITERATURE WORK

The research article [1], states that different data mining techniques could be adopted to analyze the crop yield prediction with existing data. The researchers have also addressed the problem for not using the Biclustering techniques in the field of agricultural yield prediction instead they have used K means algorithm for partitioning the samples into clusters. They have not considered the values at the end points for partition in which case Biclustering technique could be used.

[2] K-means algorithm is used for soil classification using various factors. The estimation of soil water parameter and simulation of other weather variables are done with the help of k-nearest algorithm. The application of support vector machine is the crop classification.
Soil classification has been done using Naïve Bayes classifier. K-means approach is used to analyze the crop yield.

D Ramesh et al. [3], compares the results of multiple linear regression and Density based cluster technique. The data were compared in the specific region i.e. East Godavari district of Andhra Pradesh in India. Multiple Linear Regression is applied on existing data but the results obtained are analyzed and examined using Density-based clustering technique.

Raorane et al. [4], proposes that several changes in the weather can be analyzed by Support Vector Machine (SVM is capable of classifying data samples in two disjoint clusters) and also K-means method is used to forward the pollution in atmosphere. Data mining techniques are used to monitor the wine fermentation.

In the research article [5], the authors express that government should encourage the use of data mining techniques to increase the yield and support the farmers to get reliable and timely information on crop area, crop production and land use.
It is of great importance to farmers and policy makers to have information about the yield for efficient agricultural development and for taking decisions on important issues and many other related issues to compete in the vend of crop pattern.

In the research article [6], Data mining plays a crucial role to make decisions on several issues that are related to the agriculture field. Crop yield predicting is helpful for farmers to make better policies.

In the research article [7], An attempt to predict crop yield prediction at various places of Andhra Pradesh was made. Though there are different clustering techniques, Density based clustering model is told preferable in the prediction of crop yield for approximate prediction.

IV. CONCLUSION AND FUTURE WORK

Data mining being a boon in the emerging field of technology can be used for the crop yield prediction. Different techniques of classification and clustering are used to analyze the data. We have made an attempt to understand various techniques that are used for this purpose. Biclustering techniques are rarely used in this field but having a major scope to analyze when compared to other clustering techniques. So the future enhancement would be to study Biclustering technique and use it wisely when necessary.

ACKNOWLEDGEMENT

We would like to show our gratitude to **Dr. Sr. Arpana**, Principal of Mount Carmel College for giving us an opportunity to expose ourselves into such research. We are also immensely grateful to **Regina. L. Suganthi**, HOD of Computer Science in Mount Carmel College for encouraging us to explore in the field of data mining.

REFERENCES

[1] "Data Mining Techniques and Applications to Agricultural Yield Data" by D Ramesh, B Vishnu Vardhan, Associate Professor, Department of CSE, JNTUH College of Engineering, Telangana State, India, Professor, Department of CSE, JNTUH College of Engineering, Telangana State, India.Vol 2, Issue 9, September 2013.

[2] "Analysis of Data Mining Techniques for Agriculture Data" E.Manjula, S.Djodiltachoumy, Vol.4, Issue.2, Page.1311-1313, (2016).

[3] "Analysis Of Crop Yield Prediction Using Data Mining Techniques", D Ramesh, B Vishnu Vardhan, Associate Professor, Department of CSE, JNTUH College of Engineering, Telangana State, India Professor, Department of CSE, JNTUH College of Engineering, Telangana State, India. Volume: 04 Issue: 01| Jan-2015.

[4] "Data Mining: An effective tool for yield estimation in the agricultural sector" Raorane A.A.-Department of computer science, Vivekanand College, Tarabai park Kolhapur INDIA. Kulkarni R.V.-Head of the Department, Chh. Shahu Institute of business Education and Research Centre Kolhapur 416006 INDIA, Volume 1, Issue 2, July – August 2012.

[5] "Agriculture Crop Pattern Using Data Mining Techniques" G. Nasrin Fathima, Research Scholar, Dept. of computer Science, Jamal Mohamed College, Trichirapalli, TN, India, R. Geetha , Assistant Professor , Dept. of computer Science, Jamal Mohamed College, Trichirapalli, TN, India, Volume 4, Issue 5, May 2014.

[6] "Data Mining Technique to Predict Annual Yield for Major Crops", Rajshekhar Borat, Rahul Ombale, Sagar Ahire, Manoj Dhawade, P.S. Kulkarni, Department of Computer Engineering , NBN Sinhgad School of Engineering, Pune-411041, Vol. 4, Issue 03, 2016.

[7] "Density Based Clustering Technique on Crop Yield Prediction", B Vishnu Vardhan and D. Ramesh, JNTUHCollegeofEngineering,Nachupalli,KarimnagarDist.,Andhra Pradesh,India.O Subhash Chander Goud, NIZAMCollege,Hyderabad,AndhraPradesh, India.Vol.2, No.1, March, 2014.

[8] "An Approach for Mining Accumulated Crop Cultivation Problems and their Solutions" Samhaa R. El-Beltagy, Ahmed Rafea , Said Mabrouk and Mahmoud Rafea".Cairo University, The American University in Cairo, The Central Lab for Agricultural Expert

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

IJARCCE

International Journal of Advanced Research in Computer and Communication Engineering

ICRITCSA

M S Ramaiah Institute of Technology, Bangalore

Vol. 5, Special Issue 2, October 2016

SystemsFaculty of Computers and Information, 5 Dr. Ahmed Zewail Street, 12613, Orman, Giza, Egypt, Computer Science Department, AUC Avenue, P.O. Box 74, New Cairo 11835, Egypt.Ministry of Agriculture and Land Reclamation, Giza, Egypt. 2009

[9] "Agriculture Crop Pattern Using Data Mining Techniques" G.Nasrin Fathima, Research Scholar, Dept. of computer Science,Jamal Mohamed College, Trichirapalli, TN, India. R. Geetha , Assistant Professor , Dept. of computer Science, Jamal Mohamed College, Trichirapalli, TN, India. Volume 4, Issue 5, May 2014

[10] "A survey on Data Mining Techniques for Crop Yield Prediction" Ramesh A. Medar Dept. of Computer Science & Engineering Gogte Institute of Technology Belgaum, Karnataka, India Vijay. S. Rajpurohit Dept. of Computer Science & Engineering Gogte Institute of Technology Belgaum, Karnataka, India. Volume 2, Issue 9, September 2014

[11] "Data mining Techniques for Predicting Crop Productivity – A review article"1S.Veenadhari,2 Dr. Bharat Misra, 3Dr. CD Singh1,2 Mahatma Gandhi Gramodaya Vishwavidyalaya, Chitrakoot, Satna, India 3Central Institute of Agricultural Engineering, Bhopal, India. IJCST Vol. 2, Issue 1, March 2011.