IJARCCE

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

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 5, Issue 2, February 2016

Brief Survey of data mining Techniques Applied to applications of Agriculture

Ami Mistry[1], Vinita Shah[2]

Research Scholar, Department of Information Technology, G.H. Patel collage of Engineering & Technology, Vallabh Vidyanagar[1]

Assistant Professor, Department of Information Technology, G.H. Patel collage of Engineering & Technology, Vallabh Vidyanagar[2]

Abstract: Survey made on this area reveals the importance of data mining techniques on agriculture. Lots of data mining Techniques have been used in agriculture [2]. We present some of the most used data mining techniques in the field of agriculture [1]. In the near future the penetration of Information Technology and Agriculture results is more interesting area of research. The main aim of the work is to improve and substantiate the validity of yield prediction which is useful for the farmers [6]. Agricultural crop production depends on various factors such as biology, climate, economy and geography. Several factors have different impacts on agriculture, which can be quantified using appropriate statistical methodologies. Agronomic traits such as yield can be affected by a large number of variables. In this survey, we analyzed a DM methods like clustering, classification models to select the most relevant method for the prospect [32].

Keywords: Agriculture, Yield Prediction, agricultural productivity, Classification, Clustering.

I. INTRODUCTION

Data mining techniques are widely used in various sectors of the economy. They were initially used by large companies to analyze consumer data for different perspectives. Data were then analyzed and useful information was extracted to achieve increased profitability.
In the subsequent time, Data Mining techniques were penetrated into various other sectors such as Marketing, Business, Medical and Agriculture [6].

Crop yield prediction is an important area of research which helps in ensuring food security all around the world [7].Agriculture is the backbone of Indian Economy. In India, majority of the farmers are not getting the expected crop yield due to several reasons [5]. The agricultural yield is primarily depends on weather conditions. Understanding the relative Importance of these Climate factors to crop yield variation could provide valuable information about crop planting and management under climate change condition for policymakers and farmers.

The volume of data is enormous in Indian agriculture. The data when become information is highly useful for many purposes [5]. India still depends solely on monsoon rainfall. The climate variations need to be addressed and an analysis is to be made in order to help the farmers to maximize the crop productivity [6].The hardware and the software related Data Mining and Ware Housing tools are useful to extract the knowledge from huge databases and the statistical methods are used to predict the future crop productivity[6].

For research, we have considered the effects of environmental (weather), biotic (pH, soil salinity) and area of production as factors towards crop production in any place. Taking these factors into consideration as datasets for various districts, we applied suitable data mining techniques to obtain crop yield predictions [7].

II. DATA MINING TECHNIQUES

Data Mining is the discovery for knowledge of analyzing enormous set of data by extracting the meaningful data and thereby predicting the future trends with them. The available data needs to be turned into useful information in the field of Information Technology. This useful information is further used for various applications.

Data Mining deals with what kind of patterns can be mined. Based on the kind of data to be mined, there are two kinds of functions involved in Data Mining: Descriptive model and Predictive model. The Descriptive model identifies patterns or relationships in data and deals with general properties of data in the database. The predictive model is the process of finding a model which describes the data classes or concepts, the purpose being to be able to use this model to predict the class of objects whose class label is unknown[6].

Data mining techniques are mainly divided into two groups, viz. classification and clustering techniques. Classification techniques are designed for classifying unknown samples using information provided by a set of classified samples. This set is usually referred to as a training set, because, it is generally used, to train the classification technique i.e. how to perform its classification. If a training set is not available, there is no previous knowledge about the data to classify. In this case, clustering techniques were used to split a set of unknown samples into clusters [6].

DOI 10.17148/IJARCCE.2016.5263

III. METHODS

A. Classification Techniques:

i) Linear Regression
It is a statistical measure that can be used to determine the strength of the relationship between one dependent variable and a series of other changing variables known as independent variables (regular attributes). If independent variable contains multiple input attributes like in our research (rainfall, sunshine hours, humidity, pH etc), then it is termed as multiple linear regressions. Linear regression provides a model for the relationship between a scalar variable and one or more explanatory variables. This is done by fitting a linear equation to the observed data [31].

ii) Artificial neural network
One widely used artificial neural network, back-propagation neural network (BPNN), was applied to predict rice yield because of its simplicity in structure and robustness in simulation of nonlinear systems [18]. A typical three-layer BPNN comprising one input layer, a hidden layer, and an output layer were used in the current study (Fig. 1). The neurons of adjacent layers are connected by the nodes' weights. There are two weights in the three-layer BPNN, which are vij between input and hidden layers and ujk between hidden and output layers. The aim of BPNN is to constantly modify the weights of connections between contiguous layers based on the deviation between actual values and outputs until the accuracy of the model meets the requirement of forecasting. The BPNN model can be used to forecast with new data when the weights are determined after numerous modifications [4]. The Levenberg-Marquardt algorithm [19] combined with Newtonian gradient descent algorithm was used to adjust the connection weights and biases to minimize the error. The number of neurons in the hidden layer was usually determined by trial and error [20].

iii) k-Nearest Neighbor (KNN)
The k-nearest neighbor algorithm compares a given test example with training examples which are similar. Each example denotes a point in an n dimensional space. Thus, all of the training examples are saved in an n-dimensional pattern space. K is a positive integer, usually small. For our purpose, the basic k-NN algorithm was applied.

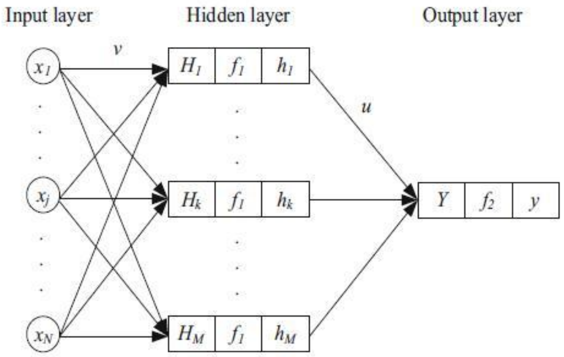

Fig1. The structure of a typical three-layer artificial neural network model

It first finds the k examples from the training set that are closest to the unknown example. Then it takes the most common occurring classification for the k examples [31].

iv) Regression Tree
In Fig.2 the red portion represents region 2 while the white portion represents region 1.The corresponding training, testing and cross-validation error for each crop type for the corresponding regions are calculated. For training sample, 70% of the data is taken, for testing 30% of the data is taken, and for cross-validation 10 partitions of the training set is done to find the error [9].

v) Support Vector Machine
The current study investigated the applicability of support vector machines (SVMs) in determining the relative importance of climate factors (mean temperature, rainfall, relative humidity, sunshine hours, daily temperature range, and rainy days) to yield variation of paddy rice in southwestern China.[4] Support vector machine (SVM) which was originally developed by Vapnik (1998) has been widely applied to many different fields, such as signal process and time series analysis. Based on the statistical learning theoryand Structural risk minimization principle, SVM is less Vulnerable to over fitting problem and it uses a hypothesis space of linear functions in a higher dimensional feature space [14]. Studies have demonstrated that SVMs are superior to traditional artificial neural networks in solving classification and regression problems due to their good generalization ability [14; 15; 16; 17].

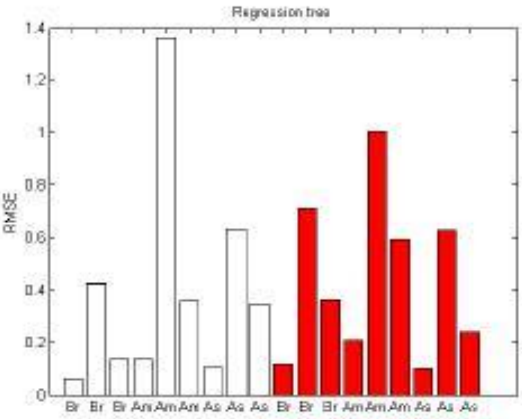

Fig2.Error found from regression tree

Nonlinear regression, Ensemble learning can also be used as a classifier.

B. Clustering Techniques:

i) K-means clustering
The K-means clustering algorithm is used to produce nonhierarchical groups of similar points in the data based on the centroid. For particular research, k-means clustering was used upon the selected districts according to the categorized types what mentioned previously [7].

ii) SOM
In clustering self-organizing maps (SOM) was used over K Means to cluster yearly average dataset. One benefit of

DOI 10.17148/IJARCCE.2016.5263

SOM over K-Means is that SOM has low classification error than KMeans [34]. This enables it to classify data points in the correct cluster. Another benefit of SOM is that it reduces the dimension of its input data points while K-Means fails to make it. Also several studies suggest that K-Means is more prone to perturbations arising from noise in the dataset, while SOM is stable to noise in the dataset.

iii) Density-based Clustering
The primary idea of Density-based clustering techniques is that, for each point of a cluster, the neighborhood of a given unit distance contains at least a minimum number of points. In other words the density in the neighborhood should reach some threshold. However, this idea is based on the assumption that the clusters are in the spherical or regular shapes. These methods group the objects according to specific density objective functions. Density is usually defined as the number of objects in a particular neighborhood of data objects. In these approaches, a given cluster continues to grow as long as the number of objects in the neighborhood which exceeds some parameter. This is considered to be different from the idea in partitioning algorithms that use iterative relocation of points that give a certain number of clusters [5].

iv) Weight based clustering
Density-based clustering technique tries to divide the data into non-equal clusters. This mechanism is being utilized by the Weight-based clustering for the purpose of dividing the training set into non-equal clusters. The process of estimation done by Density-based clustering technique using the weights given for each parameters such as Area of Sowing, Yield and Fertilizers (Nitrogen, Phosphorous and Potassium) is

$$\text{Minimum Production} = Y_{e1}*[\,(X_{min2}*0.75)+(X_{min3}*0.70)+(X_{min4}*0.20)+(X_{min5}*0.20)+(X_{min6}*0.10)\,]$$

$$\text{Maximum Production} = Y_{e1}*[\,(X_{max2}*0.95)+(X_{max3}*0.90)+(X_{max4}*0.20)+(X_{max5}*0.20)+(X_{max6}*0.10)\,]$$

where xmin2 is area of sowing density picked up minimum distance point in the cluster and is given the weight of 0.75, xmin2 is yield given the weight of 0.70 and so on until all parameters are given weights randomly. The process is done for the purpose of minimum production and maximum production using Weight-based clustering is shown below [6].

$$\text{Minimum Production} = Y_{e1}*[\,(X_{min2}*W_2)+(X_{min3}*W_3)+(X_{min4}*W_4)+(X_{min5}*W_5)+(X_{min6}*W_6)\,]$$

$$\text{Maximum Production} = Y_{e1}*[\,(X_{max2}*W_2)+(X_{max3}*W_3)+(X_{max4}*W_4)+(X_{max5}*W_5)+(X_{max6}*W_6)\,]$$

where $W_2, W_3, W_4, W_5, \dots\dots\dots\dots, W_m$ are the weights of the respective parameters.

Similar effort has also been taken in this study [33] using the DSSAT (Decision Support System for Agro Technology Transfer) model. This model is usually used to test the relationship of rice yield with the elements present in the environment. Furthermore, there are two approaches to investigate the impact of climate change on crop production which include the crop suitability approach and the production function approach [27]. Researchers were found that the yields of winter wheat are reduced when temperatures rise, due to the consequent reduction of the growth phases of the plant [28] and also concluded that the complexity of a model was based on the level of detailed analysis [29] or it was less detailed with only estimations of moisture content [30].

IV.COMPARISION

Root Mean Square Error (RMSE) is used to describe how well a machine learning algorithm performs on a certain data set [7]. From the RMSE comparison it's shown that different model provide the better result for the different crops. ANN provides better prediction for some of the crops, which have more missing values than others, for example wheat, potato and Aus. Linear regression provide better performance of predicting boro and amon [7]. The results indicate that sunshinehours and daily temperature range play critical roles in rice yield variability in the current study area. This confirms the findings reported by previous studies [21]. For example, Liu et al [22] found that daily temperature range and sunshine hours had higher direct and indirect effects on rice yield variation using path analysis. Other studies also reported that sunshine hours had a positive correlation with rice yields, since the increasing of sunshine hours could accelerate photosynthesis and hence increase crop yield [23; 24; 25; 26]. Both rainfall and rainy days have lower influence on rice yield variation than sunshine hours and daily temperature range. The models MLR and density based clustering when compared, give us the best mechanism to utilize in order to predict the crop yield and regulate the production by controlling the independent factors [8].

V. CONCLUSION

In the future, more indicators, such as soil parameters and management practices, will be needed to investigate the variability in crop yields at regional or national scales [4]. Initially the statistical model Multiple Linear Regression technique is applied on existing data. The results so obtained were verified and analyzed using the Data Mining technique namely

Density-based clustering technique [5]. 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 [7]. This will enable to have a better predictive model with more accurate results. Also additional samples of the past years which has not included here can be used to give better results [9].

REFERENCES

[1] Mucherino ,Petraq Papajorgji , P. M. Pardalos,"A survey of data mining techniques applied to agriculture", Springer,2009.
[2] S.S. Bhaskar, L. Arokiam, V. Arul Kumar, L. Jeyassimaan, "A

DOI 10.17148/IJARCCE.2016.5263

Brief Survey Of Data Mining Techniques To agriculture Applications", Madwell Journels,2010

[3] Ramesh A. Medar,Vijay. S. Rajpurohit,"A survey on Data Mining Techniques for Crop Yield Prediction", IJARCSMS,2014

[4] Hui Chen, Wei Wu, Hong-Bin Liu, "Assessing the relative importance of climate variables to rice yield variation using support vector machines",Springer,2015.

[5] D Ramesh, B Vishnu Vardhan," Analysis Of Crop Yield Prediction Using Data Mining Techniques", IJRET, 2015.

[6] D Ramesh, B Vishnu Vardhan, "Crop Yield Prediction Using Weight Based Clustering Technique ", IJCEA, 2015.

[7] A.T.M Shakil Ahamed, Navid Tanzeem Mahmood, Nazmul Hossain, Mohammad Tanzir Kabir, Kallal Das, Faridur Rahman, Rashedur M Rahman, "Applying Data Mining Techniques to Predict Annual Yield of Major Crops and Recommend Planting Different Crops in Different Districts in Bangladesh",IEEE,2015

[8] D. Ramesh ,B Vishnu Vardhan, O Subhash Chander Goud," Density Based Clustering Technique on Crop Yield Prediction"IJEEE,2014

[9] Mohammad Motiur Rahman,Naheena Haq, Rashedur M Rahman, "Application of Data Mining Tools for Rice Yield Prediction on Clustered Regions of Bangladesh", IEEE, 2014

[10] S.Veenadhari, Dr. Bharat Misra, Dr. CD Singh," Machine learning approach for forecasting crop yield based on climatic parameters", ICCCI -2014

[11] D Ramesh , B Vishnu Vardhan," Data Mining Techniques and Applications to Agricultural Yield Data" , IJARCCE, 2013

[12] José R. Romero , Pablo F. Roncallo , Pavan C. Akkiraju , Ignacio Ponzoni , Viviana C. Echenique, Jessica A. Carballido,"Using classification algorithms for predicting durum wheat yield in the province of Buenos Aires", ELSVIER, 2013.

[13] Yunous Vagh,Jitian Xiao,"A data mining perspective of dual effect of Rainfall and Temperature on Wheat Yield", ECU, 2012.

[14] Vapnik VN (1998) Statistical learning theory. Adaptive and learning systems for signal processing, communications, and control. Wiley, New York

[15] Schölkopf B, Smola A (2002) Learning with kernels. MIT, Cambridge

[16] Yoon H, Jun SC, Hyun Y, Bae GO, Lee KK (2011) A comparative study of artificial neural networks and support vector machines for predicting groundwater levels in a coastal aquifer. J Hydrol 396:128–138 Szuster BW, Chen Q, Borger M (2011) A comparison of classification techniques to support land cover and land use analysis in tropical coastal zones. Appl Geogr 31:525– 532

[17] Jain AK, Mao J, Mohiuddin KM (1996) Artificial neural networks: a tutorial. Computer 29(3):31–44

[18] Marquardt DW (1963) An algorithm for least-squares estimation of nonlinear parameters. J Soc Ind Appl Math 11:431–441
Kanungo DP, AroraMK, Sarkar S,Gupta RP (2006) Acomparative study of conventional, ANN black box, fuzzy and combined neural and fuzzy weighting procedures for landslide susceptibility zonation in Darjeeling Himalayas. Eng Geol 85:347–366

[21] Dobermann A,Dawe D, Roetter RP, CassmanKG(2000) Reversal of rice yield decline in a long-term continuous cropping experiment. Agron J 92:633–643

[22] Liu HB, Gou Y, Wang HY, Li HM, Wu W (2014) Temporal changes in climatic variables and their impact on crop yields in southwestern China. Int J Biometeorol 58:1021–1030

[23] Evans LT (1993) Crop evolution, adaptation and yield. Cambridge University Press, Cambridge

[24] YangW, Peng S, Laza RC, VisperasRM, Dionisio-SeseML(2008) Yield gap analysis between dry and wet season rice crop growth under high-yielding management conditions. Agron J 100:1390–1395

[25] Dobermann A,Dawe D, Roetter RP, CassmanKG(2000) Reversal of rice yield decline in a long-term continuous cropping experiment. Agron J 92:633–643

[26] Zhang TY, Zhu J,Wassmann R (2010) Responses of rice yields to recent climate change in China: an empirical assessment based on longterm observations at different spatial scales (1981–2005). Agric For Meteorol 150:1128– 1137

[27] M J Foulkes, "Raising Yield Potential of Wheat", Journal of Experimental Botany, vol. 62, 2011, pages: 469-486.

[28] G R Batts, "Effects Of CO2 And Temperature on Growth and Yield of Crops of Winter Wheat over Four Seasons", European Journal of Agronomy, vol. 7, 1997, pages: 43-52.

[29] R J Brooks, "Simplifying Sirus : Sensitivity Analysis and Development of A Meta-Model for Wheat Yield Prediction", European Journal of Agronomy, vol. 14, 2001, pages : 43-60.

[30] R V Martin, "Seasonal Maize Forecasting for South Africa and Zimbabwe Derived From an Agro climatological Model", Journal of Applicable Meteorology, vol. 39, 2000, pages : 1473-1479.

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

[32] Avat Shekoofa1, Yahya Emam2, Navid Shekoufa3, Mansour Ebrahimi4, Esmaeil Ebrahimie2,5*," Determining the Most Important Physiological and Agronomic Traits Contributing to Maize Grain Yield through Machine Learning Algorithms: A New Avenue in Intelligent Agriculture",2014.

[33] Jayanta Kumar Basak. "Climate Change Impacts on Rice Production in Bangladesh: Results from a Model", [Online].Available: http://www.unnayan.org/ documents/ Climatechange/climchange_impacts_rice_ production.pdf

[34] Fernando Bacao, et al., "Self-organizing Maps as Substitutes for KMeans Clustering," Springer Computational Science – ICCS 2005 Lecture Notes in Computer Science, vol 3516,pp 476-483,2005

DOI 10.17148/IJARCCE.2016.5263