A Model for Prediction of Crop Yield

E. Manjula
Pachaiyappas College
India
manjulababu123@gmail.com

S. Djodiltachoumy
Pachaiyappas College
India
djodilatchoumy@hotmail.com

Abstract- **Data Mining is emerging research field in crop yield analysis. Yield prediction is a very important issue in agricultural. Any farmer is interested in knowing how much yield he is about to expect. In the past, yield prediction was performed by considering farmer's experience on particular field and crop. The yield prediction is a major issue that remains to be solved based on available data. Data mining techniques are the better choice for this purpose. Different Data Mining techniques are used and evaluated in agriculture for estimating the future year's crop production. This research proposes and implements a system to predict crop yield from previous data. This is achieved by applying association rule mining on agriculture data. This research focuses on creation of a prediction model which may be used to future prediction of crop yield. This paper presents a brief analysis of crop yield prediction using data mining technique based on association rules for the selected region i.e. district of Tamil Nadu in India. The experimental results shows that the proposed work efficiently predict the crop yield production.**

Keywords - Agriculture, Yield Prediction, Data Mining

1. INTRODUCTION

From ancient period, agriculture is considered as the main and the foremost culture practiced in India. Ancient people cultivate the crops in their own land and so they have been accommodated to their needs. Therefore, the natural crops are cultivated and have been used by many creatures such as human beings, animals and birds. The greenish goods produced in the land which have been taken by the creature leads to a healthy and welfare life. Since the invention of new innovative technologies and techniques the agriculture field is slowly degrading. Due to these, abundant invention people are been concentrated on cultivating artificial products that is hybrid products where there leads to an unhealthy life. Nowadays, modern people don't have awareness about the cultivation of the crops in a right time and at a right place. Because of these cultivating techniques the seasonal climatic conditions are also being changed against the fundamental assets like soil, water and air which lead to insecurity of food.

By analyzing all these issues and problems like weather, temperature and several factors, there is no proper solution and technologies to overcome the situation faced by us. In India there are several ways to increase the economical growth in the field of agriculture. There are multiple ways to increase and improve the crop yield and the quality of the crops. Data mining also useful for predicting the crop yield production.

Generally, data mining is the process of analyzing data from different perspectives and summarizing it into useful information. Data mining software is an analytical tool that allows users to analyze data from many different dimensions or angles, categorize, and summarize the relationships identified. Technically, data mining is the process of finding correlations or patterns among dozens of fields in large relational databases. The patterns, associations, or relationships among all this data can provide information. Information can be converted into knowledge about historical patterns and future trends. For example, summary information about crop production can help the farmers identify the crop losses and prevent it in future.

Crop yield prediction is an important agricultural problem. Each and Every farmer is always tries to know, how much yield will get from his expectation. In the past, yield prediction was calculated by analyzing farmer's previous experience on a particular crop. The Agricultural yield is primarily depends on weather conditions, pests and planning of harvest operation. Accurate information about history of crop yield is an important thing for making decisions related to agricultural risk management. This research focuses on

evolution of a prediction model which may be used to predict crop yield production. The proposed method use data mining technique to predict the crop yield production based on the association rules.

2. BACKGROUND

2.1. Applications in Agriculture

There are several applications in the field of agriculture. Some of them are listed below.

2.1.1 Crop Selection and Crop Yield Prediction

To maximize the crop yield, selection of the appropriate crop that will be sown plays a vital role. It depends on various factors like the type of soil and its composition, climate, geography of the region, crop yield, market prices etc. Techniques like Artificial neural networks, K-nearest neighbors and Decision Trees have carved a niche for themselves in the context of crop selection which is based on various factors. Crop selection based on the effect of natural calamities like famines has been done based on machine learning (Washington Okori, 2011). The use of artificial neural networks to choose the crops based on soil and climate has been shown by researchers (Obua, 2011). A plant nutrient management system has been proposed based on machine learning methods to meet the needs of soil, maintain its fertility levels, and hence improve the crop yield (Shivnath Ghosh, 2014). A crop selection method called CSM has been proposed which helps in crop selection based on its yield prediction and other factors (Kumar, 2009).

2.1.2 Weather Forecasting

Indian agriculture mainly relies on seasonal rains for irrigation. Therefore, an accurate forecast of weather can reduce the enormous toil faced by farmers in India including crop selection, watering and harvesting. As the farmers have poor access to the Internet as a result of digital-divide, they have to rely on the little information available regarding weather reports. Up-to-date as well as accurate weather information is still not available as the weather changes dynamically over time. Researchers have been working on improving the accuracy of weather predictions by using a variety of algorithms. Artificial Neural networks have been adopted extensively for this purpose. Likewise, weather prediction based on machine learning technique called Support Vector Machines had been proposed (M.Shashi, 2009). These algorithms have shown better results over the conventional algorithms.

2.1.3 Smart Irrigation System

Farming sector consumes a huge portion of water in India. The levels of ground water are dropping down day-by-day and global warming has resulted in climate changes. The river water for irrigation is a big issue of dispute among many states in India. To combat the scarcity of water, many companies have come up with sensor based technology for smart farming which uses sensors to monitor the water level, nutrient content, weather forecast reports and soil temperature. EDYN Garden sensor is another example (Gupta, 2016). However, the high cost of such devices deters the small land owners and farmers in India to use them. These smart devices are being designed on the principles of machine learning. The nutrient content of soil can also be recorded using the sensors and hence used for supplying fertilizers to the soil using smart irrigation systems. This will also reduce the labor cost in the fields, which is a huge crisis being faced by the Indian farmers these days.

2.2. Related Work

Agricultural management needs simple and accurate estimation techniques to predict rice yields in the planning process (Ji & Wan, 2007). The necessity of the present study were to: (Washington Okori, 2011) identify whether artificial neural network (ANN) models could effectively predict rice yield for typical climatic conditions of the mountainous region, (Miss.Snehal, 2014) evaluate ANN model performance relative to variations of developmental parameters and (Shivnath Ghosh, 2014) compare the effectiveness of multiple linear regression models with ANN models.

Maize crop forecasting has been done using multilayered feed forward network of ANN (Prajneshu, 2008). They considered maize crop yield data as response variable and total human labour, farm power, fertilizer consumption, and pesticide consumption as predictors.

Generalized Regression Neural Networks (GRNN) method is used for forecasting of agricultural crop production (Chaochong, 2008). They found GRNN to be a good technique for prediction grain production in rural areas. It was reported that GRNN model is suitable for non-linear, multi-objectives and multivariate forecasting.

Crop prediction model framework was developed and concluded that climate related variables were not the main determinants of corn yield, rather yield was greatly affected by planting practices, particularly by the application right amount of fertilization (Rossana MC, 2013).

Evaluation of modified k-Means clustering algorithm in crop prediction is demonstrated by the researcher (Utkarsha P, 2014). Their results and evaluation showed the comparison of modified k-Means over k-Means and-Means++ clustering algorithm and found that the modified k-Means has achieved the maximum number of high quality clusters, correct prediction of crop and maximum accuracy count.

A model was developed for forecasting the yield of the sugarcane in Coimbatore district by using the fortnightly weather variable such as average daily maximum and minimum temperature, relative humidity in the morning and evening and total fortnightly rainfall and the yield data (Priya SRK, 2009).

Time series analysis is a method to analyze time on parametric, series data to extract meaningful statistics and other characteristics of the data. Time series forecasting is a model to predict future values based on previously observed values. New concept of crop yield under average climate conditions was described and it is used in time series techniques on the past yield data to set up a forecasting model.

3. CROP YIELD PREDICTION

Data Mining is widely applied to agricultural issues. Data Mining is used to analyze large data sets and establish useful classifications and patters in the data sets. The overall goal of the Data Mining process is to extract the information from a data set and transform it into understandable structure for further use.

This paper analyzes the crop yield production based on available data. The Data mining technique was used to predict the crop yield for maximizing the crop productivity. Figure 1 shows the flow of proposed crop yield prediction.

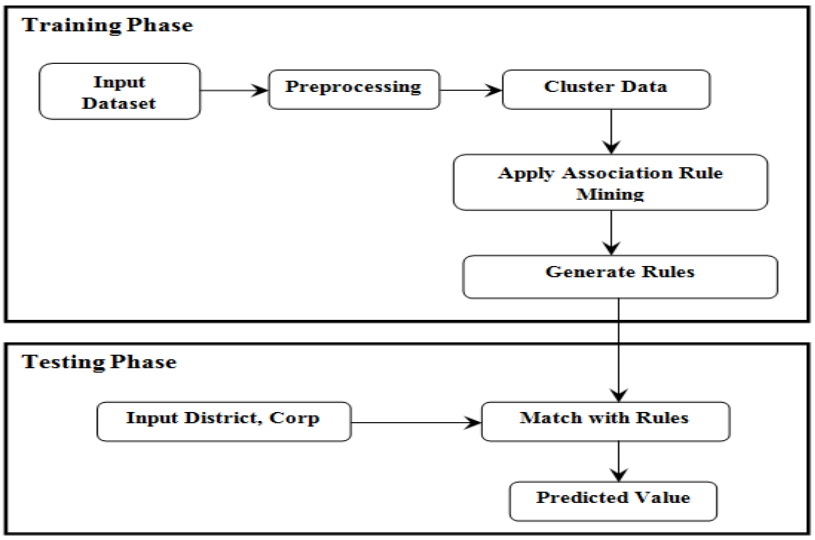

Figure 1. Proposed Work Flow

3.1 Overview of Dataset

To start with any data mining problem, it is first necessary to bring all the data together. The data used for this proposed work are obtained for the years from 2000 to 2012 for district of Tamil Nadu in India. The preliminary data collection is carried out for districts of Tamil Nadu in India. Each area in this collection is identified by the respective longitude and latitude of the region.

The data are taken in nine input variables. The variables are Year, District, Crop, Area, Tanks, Bore Wells, Open Wells, Production and Yield. Table 1 shows the description of input variables.

Table : 1 Data Set Description

Variable	Description
Year	The data was taken from the year 2000-2012
District	The data was collected 31 district of Tamil Nadu in India
Crop	Plants like a Paddy, Cholam, Cumbu and Ragi etc. from particular area
Area	The total area of agriculture plants region in Hectares.
Tanks	No of Tanks in a particular district
Bore Wells	No of Bore Wells in particular district
Open Wells	(Sole Irrigation) No of Open Wells in particular district
Production	The production of crop in the specified year in Metric Tons
Yield	It specifies in Kilogram per hectare

3.2 Methodology

The proposed methodology contains two phases: Training Phase and Test Phase. In the training phase the data was collected and preprocessed. The pre-processed data was clustered using k-means clustering algorithm. The association rule mining process will apply on clustered data to find the rules. The training phase ends with number of generated rules. In the testing phase, the yield value is predicted based on the generated rules.

The work starts with preprocessing step. In this step the collected data was pre-processed. In the pre-processing, some data was removed from the data set. Some of the area was not suitable for crop production. So that data will be removed.

3.3 Data Clustering

In the clustering step, the preprocessed data was clustered using k-means clustering algorithm. K-means is the popular clustering algorithm in this category. The k value refers to the number of clusters in which the data are partitioned. Clusters are represented by their centers. The basic idea is that each sample should be closer to the center of its own cluster. If this is not verified, then the partition is modified, until each sample is closer to the center of the cluster it belongs to. The distance function between samples plays an important role, since a sample can migrate from a cluster to another one based on the values provided by the distance function.

3.4 Data Conversion

The clustered data was converted into 0's and 1's. The data conversion process is mainly used for association rule mining step. The data variables Area, Tanks, Bore Wells, Open Wells, Production and Yield are converted into Low, Medium and High. For each clustering result the data is converted into area Low, area Medium, area High, tank Low, tank Medium, tank High, borewell Low, borewell Medium, borewell High, open well Low, open well Medium, open well High, production Low, production Medium, production High, yield Low, yield Medium, yield High.

3.5 Association Rule Mining

Once the data has been collected and converted, association rule mining may begin. In this step rules are created using frequent pattern mining. Association rules identify the relationships among a set of items or objects in a database (Utkarsha P, 2014).

Mining association rules can be reduced to mine frequent items. Association rule mining is a two-step process:

- Find all frequent items as determined by minimum support.
- Generate strong association rules that satisfy minimum support and minimum confidence.

For frequent pattern mining most common approach, Apriori algorithm, was used to reduce the computational effort.

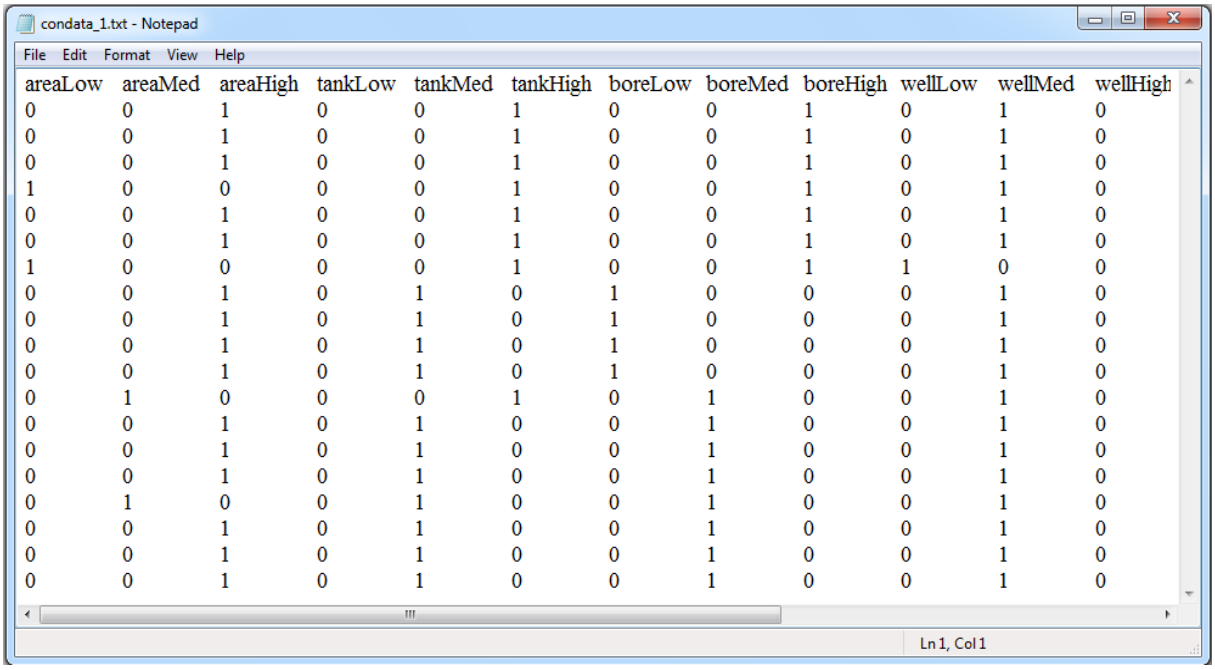

Figure 2. Converted data of Cluster-1

4. EXPERIMENTAL RESULT

The proposed work is tested on collected agriculture data obtained for the years from 2000 to 2012. The data is collected from 31 district of Tamil Nadu in India. It contains nearly 10000 records. The overview of data set is already explained Table 1 shows the description of agriculture data set. Table 2 shows the sample dataset.

Table : 2 Sample Data Set

Year	District	Crop	Area	Tanks	Bore Wells	Open Wells	Production	Yield
2000	Kancheepuram	Paddy	132855	77953	10268	46447	435500	3278
2001	Tiruchirapalli	Cholam	33672	4946	4118	44054	19290	573
2002	Villupuram	Cumbu	27294	14246	38473	75241	10774	395
2003	Thiruvannamalai	Ragi	4495	10231	147	82026	6907	1537
2004	Salem	Maize	13785	236	9103	70065	25639	1860
2005	Coimbatore	RedGram	365	555	19608	96357	197	540
2006	Perambalur	Chilies	2133	7053	18807	31768	3605	1690

The collected data set was pre-processed and some records are removed from the data. Now the data set contains nearly 7000 records. After pre-processing step, the data set was clustered based on the water level i.e Tank, Bore Well and Open Well using k-means clustering algorithm. Here the k value is set as four. The data set is grouped in to four clusters. Table 3 shows the cluster no with count of records in each cluster.

Table : 3 Cluster Info

Cluster-Id	No of Records
1	1770
2	662
3	3309
4	1305

The data in each cluster is converted into binary values. Figure 2 shows the converted data of cluster-1. For each cluster the frequent pattern is mined with minimum support of 0.3. Table4 shows the no of frequent pattern in each cluster.

Table : 4 Frequent Pattern in Each Cluster

Cluster-Id	Number of Frequent Pattern	Average Support
1	64	0.691
2	72	0.558
3	87	0.573
4	87	0.474

Table : 5 Frequent Pattern with High Support

Frequent_n	Frequent Pattern	Support
1	areaMed	0.993
2	areaMed, proMed	0.982
3	areaMed, proMed, yieMed	0.963
4	areaMed, boreMed, proMed, yieMed	0.816
5	areaMed, tankMed, boreMed, proMed, yieMed	0.681
6	areaMed, tankMed, boreMed, wellMed, proMed, yieMed	0.549

Table 5 shows high support with frequent pattern. There are 151 distinct frequent pattern is identified from all clusters and these patterns generate 193 distinct association rules with minimum support of 0.3 and minimum confidence of 0.7.

Table : 6 Rules

Rule No	Rule
1	areaMed,boreMed,proMed ⇒yieMed
2	areaMed,boreMed ⇒yieMed
3	areaMed,proMed ⇒yieMed
4	tankMed,boreHigh⇒ yieMed
5	areaMed,tankMed,boreHigh,wellMed⇒yieMed
6	tankMed,boreMed,wellHigh ⇒ yieMed
7	areaMed,tankMed,wellMed,proMed ⇒ yieMed

Figure 2. Overall Accuracy and Error Rate

The crop yield was predicted based on the generated rules. The district and crop is input to the prediction model. Figure 1 shows the overall accuracy and error rate of prediction result.

5. CONCLUSION

Crop yield prediction is still remaining as a challenging issue for farmers. The aim of this research is to propose and implement a rule based system to predict the crop yield production from the collection of past data. This has been achieved by applying association rule mining on agriculture data from 2000 to 2012.

REFERENCES

Chaochong, J. M. (2008). Forecasting Agricultural Production via Generalized Regression Neural Network. IEEE .

Gupta, A. S. (2016). Need Of Smart Water Systems In India. International Journal of Applied Engineering Research, 2216-2223.

Hong-Ying L, Y.-L. H.-J.-M. (2012). Z.Crop yield forecasted model based on time series techniques. Journal of Northeast Agricultural University (English Edition).

J. Khazaei, M. R. (2008). Yield estimation and clustering of chickpea genotypes using soft computing techniques. Agron J 2008.

Ji, B. S., & Wan, J. (2007). Artificial neutral networks for rice yield prediction in mountainous regions. Joural of Agricultural science.

Kumar, R. M. (2009). Crop Selection Method to Maximize Crop Yield Rate Using Machine Learning Technique. International Conference on Smart Technologies and Management for Computing, Communication, Controls, Energy and Materials, 1 (1).

M.Shashi, Y. a. (2009). Atmospheric Temperature Prediction using Support Vector Machines. International Journal of Computer Theory and Engineering, 1 (1).

Miss.Snehal, S. D. (2014). Agricultural Crop Yield Prediction Using Artificial. International Journal of Innovative Research in Electrical,Electronic, 1 (1).

Obua, W. O. (2011). Machine Learning Classification Technique for Famine Prediction. Proceedings of the World Congress on Engineering, 2.

Prajneshu, R. K. (2008). Artificial Neural Network Methodology for Modelling and Forecasting. Agricultural Economics Research Review, 21, 5-10.

Priya SRK, S. K. (2009). A study on pre-harvest forecast of sugarcane yield using climatic variables,Statistics and Applications. 1-8.

R. Agrawal, R. J. (1999). Mining the most interesting rules, in: Proceedings of the 1999. SIGKDD International Conference on Knowledge Discovery and Data Mining, 145-154.

Rossana MC, L. D. (2013). A Prediction Model Framework for Crop Yield Prediction. Asia Pacific Industrial Engineering and Management Society Conference Proceedings Cebu, Phillipines, 185.

Shivnath Ghosh, S. K. (2014). Machine Learning for Soil Fertility and Plant Nutrient Management. International Journal on Recent and Innovation Trends in Computing, 2 (2), 292-297.

Utkarsha P, N. N. (2014). Evaluation of Modified K-Means Clustering Algorithm in Crop prediction. International Journal of Advanced Computer Research., 1–1.

Washington Okori, J. O. (2011). Machine Learning Classification Technique for Famine Prediction. Proceedings of the World Congress on Engineering 2011, vol 2.