Paddy Yield Predictor Using Temperature, Rainfall, Soil pH, and Nitrogen

Chapter · January 2019

DOI: 10.1007/978-981-13-5802-9_23

Prathibha R J
Sri Jayachamarajendra College of Engineering
8 PUBLICATIONS **13** CITATIONS

Paddy Yield Predictor Using Temperature, Rainfall, Soil pH, and Nitrogen

Pooja R. Rao, Sanju P. Gowda and R. J. Prathibha

Abstract Agriculture is the backbone of India which indirectly contributes for an Indian economy. Farmers who are the drivers of agriculture are facing lot of problems for proper identification of the crops that can be cultivated for the specific soil conditions and to maximize the crops yield. All these problems are due to lack of technology and scientific techniques being used in agriculture. Crop yield varies as a result of variations in atmospheric and soil conditions. Data mining mainly focuses on methods to elicit useful knowledge from the dataset. There are several data mining approaches that can be used for the purpose of predicting crops yield and finding association among attributes contributing for the crops yield. This paper mainly intensifies on various association algorithms, namely Apriori, Eclat, and AprioriTid to find the association among temperature, rainfall, soil pH, soil nitrogen, and paddy yield.

Keywords Data mining · Association algorithms · Apriori · Eclat · AprioriTid

1 Introduction

Agriculture in India is popular for its vast diversity. It depends on climatic and weather conditions, soil components, manures used, resources available, and political and socio-economical factors. Due to variations in these factors, crops yield also gets

P. R. Rao (✉) · S. P. Gowda
Department of Information Science and Engineering, JSS Science and Technology University, Mysuru 570006, Karnataka, India
e-mail: rao.pooja32@gmail.com

S. P. Gowda
e-mail: sanju.p@hotmail.com

R. J. Prathibha
Department of Information Science and Engineering, SJ College of Engineering, Mysuru 570006, Karnataka, India
e-mail: rjprathibha@sjce.ac.in

V. Sridhar et al. (eds.), *Emerging Research in Electronics, Computer Science and Technology*, Lecture Notes in Electrical Engineering 545,
https://doi.org/10.1007/978-981-13-5802-9_23

varied. As a result of which, yield production has taken uneven path as the years rolled upon and these production changes are in accordance with different geographical locations.

Agriculture as a business is dependent on various factors which seem to be risks. These risks need to be properly managed to attain success in agriculture. To address these risks, farmers need to be guided in the right path in terms of proper identification of crops in the right season, factors influencing crops yield, required soil features for cultivation, manures to be used and so on. Yield prediction is one of the important challenges that need to be handled in agriculture. Every farmer is very much engrossed in knowing quantitative yield outcome of his crop in prior. Earlier, this was done using farmer's previous experience with the crop. Application of data mining in agriculture would help to contribute for predicting crops yield and finding association among contributing attributes based on previous historical data.

This paper mainly focuses on finding the association among temperature, rainfall, soil pH, soil nitrogen, and paddy yield. It makes use of agricultural data of Nanjangud district from past six years to find association and to predict paddy crop yield. It uses various association rule mining techniques, namely Apriori algorithm, Eclat algorithm, and AprioriTid algorithm to find association among temperature, rainfall, soil pH, soil nitrogen, and paddy yield.

This paper organization is as follows: Sect. 2 quotes the related work in the field of agriculture. Section 3 portrays the methodology and association rule mining algorithms used to build the system. Section 4 focuses on the experimentation and discussion of the results obtained. Lastly, conclusion and future enhancement are given in Sect. 5.

2 Literature Survey

Kaur et al. [1] worked on applying predictive Apriori algorithm to agricultural dataset of Patiala and Ludhiana district to analyze the effect of daily temperature, daily rainfall on paddy yield. Waikato Environment for Knowledge Analysis (WEKA) tool was used to conduct the study. Various phases involved in paddy growth were considered for the analysis. Manjula et al. [2] proposed a prediction model to predict crops yield based on the previous data available for districts of Tamil Nadu. Apriori algorithm is applied on the preprocessed dataset which was obtained using modified K-means algorithm, to predict the various crops yield. Thombare et al. [3] focused on applying K-means algorithm for reducing the input and storing the data in clusters which would result in faster search with less time. Apriori algorithm is used for predicting the crops yield. Fathima and Geetha [4] investigated the use of various data mining techniques in the field of agriculture. Apriori algorithm of association rule mining was used to find crop pattern, K-means clustering algorithm, and K-Nearest Neighbor classification algorithm were also used for the study. Khan and Singh [5] proclaimed that various association rule mining algorithms can be applied in various aspects related to agriculture. It includes the description of various association algorithms like Apriori, Pincer search algorithm, and so on. Dey et al. [6] done a survey on rice yield produced per year using SVM regression, multiple linear regression, and AdaBoost techniques by considering the attributes Rainfall, Temperature, Humidity, Area, and Yield. Ramesh and Vishnu Vardhan [7] presents a brief analysis on crop yield prediction using multiple linear regression technique applied on existing data and density-based clustering technique applied to verify and analyze the results obtained.

3 Methodology

The approach used by this paper is discussed in this section which includes data collection, data preprocessing, applying association rule mining algorithm, and analyzing association rules.

3.1 Data Collection

Agricultural data required for the work are collected from the year 2012 to 2017 for Nanjangud taluk which includes name of the farmer, taluk, hobli, village, year, temperature, rainfall, soil pH, soil nitrogen, and paddy yield.

3.2 Data Preprocessing

Collected data contains many attributes out of which only the relevant attributes like temperature, rainfall, soil pH, soil nitrogen and paddy yield were selected and numeric values of these attributes were discretized into low, medium, and high based on the respective threshold value specified for these attributes.

3.3 Applying Association Rule Mining Algorithm

For the preprocessed data, various association rule mining algorithms are applied to get frequent itemsets and strong association rules are generated from frequent itemsets obtained. Filtering only those rules which contain all five attributes and consequent part of the rule should contain only attribute paddy yield. Displaying only those rules with the highest confidence value for each of the discretized values of low, medium and high.

3.3.1 Apriori Approach

Apriori algorithm is used to find frequent patterns among items in the transactions stored in the database and association rule is generated using frequent itemsets. It proceeds by generating candidate itemset, by combining each item with every other item from the previous frequent itemset. It mainly works on property that "All non-empty subsets of a frequent itemset must also be frequent". Frequent itemset is generated from the previous candidate itemset by pruning those items whose support count does not satisfy minimum support threshold. Frequent itemset thus obtained contains items whose support is greater than the minimum support. It uses "bottom-up" approach where frequent itemsets are extended one at a time known as candidate itemsets. This continues until no further extensions can be done. The main drawback is it needs multiple scans of database to calculate the support of each item. From the frequent itemsets, association rule is generated.

3.3.2 Eclat Approach

Eclat algorithm is used to find frequent patterns among items in the transactions stored in the database and association rule is generated using frequent itemsets. It proceeds by generating candidate itemset by combining each item with every other item from the previous frequent itemset. It uses intersection-based approach to calculate support count of each candidate item. Frequent itemset is generated from the previous candidate itemset by pruning those items whose support does not satisfy minimum support threshold. Frequent itemset thus obtained contains items whose support is greater than the minimum support. It is more appropriate for vertical data layout. This continues until no further extensions can be done. From the frequent itemsets, association rule is generated.

3.3.3 AprioriTid Algorithm

AprioriTid algorithm is used to find frequent patterns among items in the transactions stored in the database, and association rule is generated using frequent itemsets. Just like the Apriori algorithm, AprioriTid algorithm proceeds by generating candidate itemset by combining each item with every other item from the previous frequent itemset. But the only difference is database is not referred for counting support after the first pass. Sets of candidate itemsets are generated for $K > 1$. When the given candidate K-item is not present in the transaction, then candidate itemsets will not be having any entry for that transaction. This leads to decrease in the number of transactions containing candidate itemsets when compared to database. As the value of K increases, there will be decrease in the number of entries in the set of candidate itemsets than the corresponding transactions due to decrease in the number of candidates in the transaction. Frequent itemset is generated from the previous candidate itemset by pruning those items whose support count does not satisfy minimum support threshold. Frequent itemset thus obtained contains items whose support is greater than the minimum support. This continues until no further extensions can be done. From the frequent itemsets, association rule is generated.

3.4 Analyzing Association Rules

Strong association rules obtained were analyzed to understand the correlation among temperature, rainfall, soil pH, soil nitrogen, and paddy yield.

4 Experimentation

4.1 Dataset

The attribute temperature specifies average temperature in degree celsius for Kharif season in particular year as specified by year attribute. The attribute rainfall specifies average rainfall in millimeter for Kharif season in particular year as specified by year attribute. The attribute soil pH specifies the pH value for a particular plot. The attribute soil nitrogen specifies the nitrogen value in kilogram per hectare for a particular plot. The attribute paddy yield specifies the yield obtained for the particular farmer plot in particular year, which is measured in quintal per hectare.

Agricultural data used for this work are obtained from the year 2012 to 2017 for Nanjangud taluk of Mysore district, Karnataka. It includes name of the farmer, taluk, hobli, village, year, temperature in degree celsius, rainfall in millimetre, soil pH, soil nitrogen in kilogram per hectare, and paddy yield in quintal per hectare. Sample data for the year 2012 and 2013 are given Table 1.

Table 1 Sample dataset

Name	Village	Year	Temperature	Rainfall	pH	Nitrogen	Yield
Sunandappa	Belagunda	2012	26.16	68.21	5.7	188.43	18
Mahadevayya	Nagarle	2012	26.16	68.21	7.2	127.4	20
Siddaraju	Nagarle	2012	26.16	68.21	6.4	158	20
Naganayaka	Haniyamballi	2012	26.16	68.21	6.5	110	21
Puttaraju	Saragooru	2013	25.28	105.43	5.8	146.5	18
Chandramma	Belagunda	2013	25.28	105.43	7.89	173.9	25
Aunkanayaka	Haniyamballi	2013	25.28	105.43	6.88	117.6	25
Hucchegowda	Kupparavalli	2013	25.28	105.43	6.24	103.24	26

4.2 Experimental Setup

The input data considered here are name of the farmer, taluk, hobli, village, year, temperature, rainfall, soil pH, soil nitrogen, and paddy yield. Preprocessing of the input results in selection of only five attributes, namely temperature, rainfall, soil pH, soil nitrogen, and paddy yield which are appropriate for building the system and converting numeric values of these attributes into low, medium, and high based on the respective threshold value specified for these attributes. Apriori or Eclat or AprioriTid algorithm reads the preprocessed data and processes the data to produce the association rules which are filtered and analyzed to understand the correlation among the attributes.

4.3 Result and Discussion

The result obtained using Apriori, Eclat, and AprioriTid algorithm contains list of association rules. The minimum support and minimum confidence value used for these algorithms is 0.1. On choosing these values to be minimum, it is possible to get maximum possible combinations among items in the database. This list contains the unfiltered rules for the given minimum support and minimum confidence value. The sample of unfiltered rules obtained for given minimum support and minimum confidence value of 0.1 is given in Table 2.

The list of rules thus obtained is filtered such that the filtered rule should contain all the five attributes, namely temperature, rainfall, soil pH, soil nitrogen, and paddy yield, and the consequent part of the rule should contain only paddy yield. Displaying only those rules with the highest confidence value for each of the discretized values of low, medium, and high.

The outcome obtained using Apriori algorithm for Nanjangud taluk includes a list of association rules. Association rule list contains only the filtered rules, which suggest the most suitable condition for getting the high and medium paddy yield.

Table 2 Obtained association rules

Rule X	→	Rule Y	Confidence (%)
N-Medium	→	T-High, Y-Medium	59.86
N-Medium	→	Y-Medium	59.86
N-Medium, P-Medium	→	T-High, Y-Medium	58.64
N-Medium, P-Medium	→	Y-Medium	58.23
N-Medium, P-Medium, R-High	→	T-High, Y-Medium	67.56
N-Medium, P-Medium, R-High	→	Y-Medium	56.78
N-Medium, P-Medium, R-High, T-High	→	Y-Medium	55.58
N-Medium, P-Medium, R-High, Y-Medium	→	T-High	100.00

Rules obtained define the association among attributes temperature, rainfall, soil pH, soil nitrogen, and paddy yield. The results obtained by Apriori algorithm for the given dataset are depicted in Table 3.

The outcome obtained using Eclat algorithm for Nanjangud Taluk includes a list of association rules. Association rule list contains only the filtered rules, which suggest the most suitable condition for getting the high and medium paddy yield. Rules obtained define the association among attributes temperature, rainfall, soil pH, soil nitrogen, and paddy yield. The results obtained by Eclat algorithm for the given dataset are depicted in Table 4.

The outcome obtained using AprioriTid algorithm for Nanjangud taluk includes a list of association rules. Association rule list contains only the filtered rules, which suggest the most suitable condition for getting the high and medium paddy yield. Rules obtained define the association among attributes temperature, rainfall, soil pH, soil nitrogen, and paddy yield. The results obtained by AprioriTid algorithm for the given dataset are depicted in Table 5.

The association rules obtained among the attributes temperature, rainfall, soil pH, soil nitrogen, and paddy yield for available datasets of Nanjangud taluk defines that paddy yield will be medium when temperature is high and attributes rainfall, soil pH, soil nitrogen are medium and paddy yield will be high when temperature, rainfall are high and soil pH, soil nitrogen are medium. Below table defines the maximum, average, and minimum time needed to run Apriori, Eclat, and AprioriTid algorithms applied on available 512 datasets of Nanjangud taluk. These values are noted on running Apiori, Eclat, and AprioriTid algorithms for ten times. Time comparison among Apriori, Eclat, and AprioriTid algorithm is given in Table 6.

Table 3 Filtered association rules obtained from Apriori algorithm

Rule X	→	Rule Y	Confidence (%)
N-Medium, P-Medium, R-Medium, T-High	→	Y-Medium	66.67
N-Medium, P-Medium, R-High, T-High	→	Y-High	40.17

Table 4 Filtered association rules obtained from Eclat algorithm

Rule X	→	Rule Y	Confidence (%)
N-Medium, P-Medium, R-Medium, T-High	→	Y-Medium	66.67
N-Medium, P-Medium, R-High, T-High	→	Y-High	40.17

Table 5 Filtered association rules obtained from AprioriTid algorithm

Rule X	→	Rule Y	Confidence (%)
N-Medium, P-Medium, R-Medium, T-High	→	Y-Medium	66.67
N-Medium, P-Medium, R-High, T-High	→	Y-High	40.17

Table 6 Comparison among Apriori, Eclat and AprioriTid algorithm

Algorithm	Min. time (in ms)	Average time (in ms)	Max. time (in ms)
Apriori	145	148.5	158
Eclat	63	68.9	76
AprioriTid	111	116.1	126

From the above observation, the result obtained using Apriori, Eclat, and AprioriTid algorithms are same. But the time taken by these algorithms to compute the same result is different as a result of change in the computational procedure being used by these algorithms. Eclat needs less time than AprioriTid, and AprioriTid needs less time than Apriori to compute the result.

5 Conclusion and Future Enhancement

Predicting the crops yield and finding association among contributing attributes is one of the important challenges to be handled in the field of agriculture. For this purpose, association rule mining algorithms, namely Apriori, Eclat, and AprioriTid algorithms are used in this work. Apriori needs multiple database scans. Eclat and AprioriTid algorithms scans data base to generate frequent-one itemset and proceeds further to generate consequent frequent itemsets based on itemset obtained in previous iteration. When it comes to efficiency among the three algorithms, Eclat is the first best and AprioriTid is the second best. The outcome of this work for Nanjangud taluk defines that paddy yield will be medium when temperature is high and attributes rainfall, soil pH, soil nitrogen are medium and paddy yield will be high when temperature, rainfall are high and soil pH, soil nitrogen are medium. Application of this work would definitely contribute to maximize the crops yield and to manage risks related to agriculture.

This work can be extended for predicting the type of crop that can be cultivated in a particular plot for high yield. Also this work can be enhanced to predict the better association rule by considering additional attributes like Phosphorus, Carbon, and Zinc. It can be applied to more agricultural crops of various locations. More number of association algorithms can be implemented on the agricultural dataset.

References

1. Kaur K, Attwal KS (2017) Effect of temperature and rainfall on paddy yield using data mining. In: 2017 7th international conference on cloud computing, data science & engineering-confluence. IEEE

2. Manjula E, Djodiltachoumy S (2017) A model for prediction of crop yield. Int J Comput Intell Inf 6(4)
3. Ms Thombare R, Ms Bhosale S, Mr Dhemey P, Ms Chaudhari A (2017) Crop yield prediction using big data analytics
4. Fathima GN, Geetha R (2014) Agriculture crop pattern using data mining techniques. Int J Adv Res Comput Sci Softw Eng 4(5):781–786
5. Khan F, Singh D (2014) Association rule mining in the field of agriculture: a survey. Int J Sci Res Publ 329
6. Dey UK, Masud AH, Uddin MN (2017) Rice yield prediction model using data mining. In: International conference on electrical, computer and communication engineering (ECCE). IEEE
7. Ramesh D, Vishnu Vardhan B (2015) Analysis of crop yield prediction using data mining techniques. Int J Res Eng Technol 4(1)