Soil Data Analysis and Crop Yield Prediction in Data Mining using R-Tool

K. Samundeeswari, K. Srinivasan

ABSTRACT

Background: Crop yield prediction is an important issue for the proper selection of crop for sowing. Earlier prediction of crop is done by the farmer’s experience on a particular type of field and crop. Predicting the crop is done by the farmer’s experience based on the factors like soil types, climatic condition, seasons and weather, rainfall and irrigation facilities.

Methods: Data mining techniques is the better choice for predicting the crop. 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 soil data. This is achieved by applying Decision Tree Algorithm on agricultural data. The main aim of this research is to pinpoint the accuracy of Decision Tree Algorithm and C 5.0 algorithm which is used to predict the crop yield.

Result: This paper presents a brief analysis of Crop yield prediction using data mining technique based decision tree algorithm and C5.0 algorithm for the selected region (Krishnagiri) district of Tamil Nadu in India. The experimental result shows that the proposed work efficiently to determine the accuracy of decision tree algorithm and also to predict the crop yield production using R-Tool.

Key words: Crop yield prediction, Data mining, Decision tree algorithm, R-Tool, Soil data set.

INTRODUCTION

Data Mining is a process of extracting hidden information from a database and transform it into a useful information for further use. Data Mining is used to find desired patterns from large data sets and establish useful classifications from the data sets. Data Mining is widely used in agricultural problems. The goal of the Data Mining issue is to extract the information from a large data set and transforming it into meaningful usable information for further use. Data mining is a tools used to predict crop yield and Agricultural output (Manjula et al., 2017).

In India, it is common that farmers are not getting the expected crop yield due to some reasons. The agricultural yield is mainly depends on weather conditions. In this perspective, the farmers necessarily requires an appropriate advice to predict the future crop productivity and an study is to be made in order to help the farmers to capitalize on the crop production in their crops. Agriculture forms the basis for food security and it is important to human beings. India’s agriculture is composed of many crops, but the primary food staples are rice and wheat. Agriculture is one of the main and important back-bone of Indian economy. The efficiency of agriculture depends on environmental Conditions and season. Capitulate of the agriculture, gives one of the quantifiable parameters that supply towards the real income of the country. This capitulate is calculated based on the area and production of each crop in the country.

The soil testing process starts with the collection of a soil sample from a field. The primary principle of soil testing is that a field can be sampled using chemical analysis of the soil sample will accurately reflect the field’s true nutrient status. The soil can be analyzed for various physical properties and chemical properties such as EC, pH, along with micro nutrients N, P, K and macro nutrients Zn, B, Cu, Fe (Kumar et al., 2009). The classification techniques can be applied to the soil data, the results of classification helps to determine the cultivation pattern. Classification and clustering techniques can be applied to analyze the soil properties. Soil testing is a pre-cultivation activity which gives a brief idea about soil structure and mineral compositions ratios. Soil testing must be done frequently to analyze the vitality of soils during the seasonal changes. The essential nutrients required for various crop growths can be estimated during soil testing process. The function of soil testing is high-yield farming is to establish the relative ability of a soil to provide crop nutrients during a particular growing season, to determine the needs and diagnosing problems such as extreme salinity or alkalinity. Soil testing is also used to direct nutrient management decisions (Rajeswari et al., 2016).

The main aim of this paper is to find the accuracy of decision tree algorithm and C 5.0 algorithm to predicate the crop yield based on available data. Different Data mining
techniques were used to predict the crop yield for maximizing the crop productivity. The definitive goal of data mining is prediction; data mining is the common way used in predictive data mining. Many algorithms were created to extract knowledge from large data sets (Snehal 2019). There are several different methodologies are used to approach this problem: classification, feature selection, j48, etc. Here we focus on decision tree algorithm-classification methodology. Classification techniques are designed for classifying unidentified samples using information provided by a set of classified samples.

**MATERIALS AND METHODS**

This paper uses agriculture raw data collected from district agricultural office at Krishnagiri. This raw data set is then preprocessed and analyzed by using R-tool. This paper describes about the progress of retrieving required data from given soil data sets. Decision tree algorithm are used for given soil data set with the following steps,

1. Agriculture data is used as input data for system.
2. Pre-process data is nothing but removing unwanted data or adding required information to make data mining easier.
3. This pre-processed data is used as input for further implementation.
4. Create two datasets from available data, one for training and other for testing.
5. Apply decision tree algorithm and c.50 algorithm.

The data used for this paper are obtained for the year 2017 from few Taluks of Krishnagiri district of Tamil Nadu in India. The preliminary data collection is carried out for all the Taluks of Krishnagiri district of Tamil Nadu in India. Micronutrients are essential for plant growth and they play an important role in balanced crop nutrition (Samundeeswari et al., 2017). They include copper (Cu), iron (Fe), manganese (Mn), zinc (Zn). They are as important to plant nutrition as primary and secondary macronutrients, though plants don’t require as much of them. A lack of any of one of the micronutrients in the soil can limit growth, even when all other nutrients are present in adequate amounts.

**Decision tree algorithm**

Decision Tree algorithm is one of the supervised learning algorithm, decision tree algorithm can be used for solving classification and regression problems. Decision Tree algorithm is to create a training data set model which is used to predict crop yield inferred from prior data (training data). The decision tree algorithm is used to solve the problem, by using tree representation. The primary challenge in the decision tree implementation is to identify which attributes we need to consider as the root node and each level. We have different attributes selection measure to identify the attribute which can be considered as the root note at each level.

**The popular attribute selection measures**

- Information gain
- Gini index

**C 5.0 algorithm**

C 5.0 algorithm is the advance version of C 4.5. C5.0 Decision Tree algorithm is one of the popular and effectively used classifier for crop yield in soil data set classification in present study. C5.0 is an algorithm used to generate a decision tree (Hari Ganesh et al., 2015). The decision trees generated by C 5.0 can be used for classification 5.0 builds decision trees from a set of training data using the concept of information entropy. The training data is a set of already classified samples. At each node of the tree, C 5.0 chooses the attribute of the data that most effectively splits its set of samples into subsets enriched in one class or the other. The splitting criterion is the normalized information gain (difference in entropy). The attribute with the highest normalized information gain is chosen to make the decision.

**R-programming**

R [R Core Team, 2015b] is a free software environment for statistical computing and graphics. It provides a wide variety of statistical and graphical techniques. R can be easily extended with 7324 packages available on CRAN3 (as of October 20, 2015).

To help users to find our which R packages to use, the CRAN Task Views 9 are a good guidance. They provide collections of packages for different tasks. Some Task Views related to data mining are:

- Machine Learning and Statistical Learning,
- Cluster Analysis and Finite Mixture Models
- Time Series Analysis,
- Natural Language Processing,
- Multivariate Statistics,
- Analysis of Spatial Data

**R Studio**

R Studio is an integrated development environment (IDE) for R and can run on various operating systems like Windows, Mac OS X and Linux. It is a very useful and powerful tool for R programming and therefore, readers are suggested to use R Studio. What you normally need is R Studio Desktop open source edition, which is free of charge. When R Studio is launched for the first time. There are four panels (Kanjana Devi et al., 2016):

- Source panel (top left), which shows your R source code.
  - If you cannot see the source panel, you can find it by clicking menu “File”, “New File” and then “R Script”. You can run a line or a selection of R code by clicking the “Run” button on top of source panel, or pressing “Ctrl + Enter”.
- Console panel (bottom left), which shows outputs and system messages displayed in a normal R console.
- Environment/History/Presentation panel (top right), whose three tabs show respectively all objects and function loaded in R, a history of submitted R code and Presentations generated with R.
- Files/Plots/Packages/Help/Viewer panel (bottom right), whose tabs show respectively a list of files, plots, R
Soil Data Analysis and Crop Yield Prediction in Data Mining using R-Tool

It is always a good practice to begin R programming with an RStudio project, which is a folder where to put your R code, data files and Figures. To create a new project, click the “Project” button at the top-right corner and then choose “New Project”. After that, select “create project from new directory” and then “Empty Project”. After typing a directory name, which will also be your project name, click “Create Project” to create your project folder and files. If you open an existing project, RStudio will automatically set the working directory to the project directory, which is very convenient. After that, create three folders as below:

- code, where to put your R source code;
- data, where to put your datasets;
- Figures, where to put produced diagrams.

In addition to above three folders which are useful to most projects, depending on your project and preference, you may create additional folders below:

- raw data, where to put all raw data;
- models, where to put all produced analytics models and
- Reports, where to put your analysis reports.

Coding for soil data using decision tree and c 5.0

```r
sdata <- read.csv("d:/data/sdata.csv",header=TRUE)
View(sdata) # view total data
summary(sdata) # data summary
dim(sdata) # dimension of the data
names(sdata) # display header of the data
library(Hmisc)
library(lattice)
library(survival)
library(Formula)
library(ggplot2)
range(sdata$CA) # range
mean(sdata$CA)
quantile(sdata$CA) # quartiles and percentiles
var(sdata$CA)
hist(sdata$CA)
library(magrittr) ## for pipe operations
sdata$CA %>% density() %>%
plot(main='Density of CA')
library(dplyr)
sdata2 <- sdata %>% sample_n(50)
sdata2$crop %>% table() %>% pie()
tab <- sdata2$crop %>% table()
percentages <- tab %>% prop.table() %>% round(3) * 100
txt <- paste0(names(tab), ",", percentages, ")
pie(tab, labels=txt)
cov(sdata$CA, sdata$PH) # covariance
cor(sdata$CA, sdata$PH) # correlation
boxplot(PH ~ crop, data = sdata)
with(sdata, plot(CA, MG, col = crop, pch = as.numeric(crop)))
#pairs(sdata)
library(scatterplot3d)
scatterplot3d(sdata$N, sdata$P, sdata$K)
```

# Decision tree

```r
dim(sdata)
head(sdata, 7)
set.seed(1234)
ind <- sample(2, nrow(sdata), replace=FALSE, prob=c(0.7, 0.3))
sdata.train <- sdata[ind==1,]
sdata.test <- sdata[ind==2,]
# train a decision tree
library(rpart)
myFormula <- N ~ P + CA + MG
sdata.rpart <- rpart(myFormula, data = sdata.train,
control = rpart.control(minsplit = 10))
fit <- rpart(myFormula, data = sdata.train, method = "class")
predict_unseen <- predict(fit, sdata.test, type = "class")
table.mat <- table(sdata.test$crop, predict_unseen)
table.mat
accuracy_Test <- sum(diag(table.mat)) / sum(table.mat)
print(paste('Accuracy for test', accuracy_Test))
print(sdata.rpart$cptable)
```

# C 5.0 algorithm

```r
library(C50)
# load data
# fit model
fit <- C5.0(crop~., data=sdata, trials=10)
# summarize the fit
print(fit)
# make predictions
predictions <- predict(fit, sdata)
# summarize accuracy
table(predictions, sdata$crop)
```

# Bagging CART

```r
library(ipred)
# fit model
fit <- bagging(crop~., data=sdata)
# summarize the fit
summary(fit)
# make predictions
predictions <- predict(fit, sdata, type="class")
# summarize accuracy
table(predictions, sdata$crop)
summary(predictions)
```
Soil Data Analysis and Crop Yield Prediction in Data Mining using R-Tool

RESULTS AND DISCUSSION

The Experiments are performed on the real world data obtained from the district agricultural office at Krishnagiri District in Tamil Nadu. Datasets considered in this work have sufficient amount of readings of nutrients and micronutrients taken from different taluks of Krishnagiri district. The dataset used in this experiment consists of 249 tuples with 24 attributes. The tuples of dataset thus define the availability of available nutrients (N,P,K), Ph, Ec, crop, fertilizer recommendation (N,P2O5,K2O,urea, super phosphate, muriate of potash), micronutrients contents (Fe,Zn,Mn,Cu) and Micro Nutrient Recommendation (FeSO4,ZnSO4, MnSO4,CuSO4) in soil.

The data are taken in twenty four input variables. The variables are:

- S.No
- CA - Calcium
- MG - Magnesium
- SO4 - Sulphur
- B - Boron
- PH - potential hydrogen
- EC - Electrical Conductivity
- N - Nitrogen
- P - Phosphorus
- K - Potassium
- Crop - type of crop
- P2O5 - Phosphorus Pentoxide
- K2O - Potassium Oxide
- Fe - Iron
- Zn - Zinc
- Mn - Manganese
- Cu - Copper
- FeSO4 - Iron Sulphate
- ZnSO4 - Zinc Sulphate
- CuSO4 - Copper Sulphate

Table 1: Input variables.

| S.No | CA  | MG  | SO4 | B   | PH  | Ec  | N   | P   |
|------|-----|-----|-----|-----|-----|-----|-----|-----|
|      |     |     |     |     |     |     |     |     |

Table 2: Summary of soil dataset.

| S.No | CA  | MG  | SO4 | B   | PH  | Ec  | N   | P   |
|------|-----|-----|-----|-----|-----|-----|-----|-----|
| Min. | 1   | 0   | 1.0 | 1.0 | 13.0| 1.0 | 27.0| 5.00|
| Max. | 249 | 125 | 1289 | 50 | 315.8| 1289 | 290.5| 64.0|

Above table describes about the summary of the soil dataset to find the accuracy of the algorithm, mean, median, minimum, maximum, first quartile and third quartile values are calculated for all the twenty four input variables.
to find the efficiency and deficiency of the soil it will be very useful for soil testing.

Pie chart show the pictorial representation of the crop grown and also we can easily identify which crop gives the high yield in that particular area according the available micro nutrient and macro nutrients and what are the nutrients are recommended to give the high yield of a particular crop in that particular area can be easily identified using the pie chart from the give data set.

Description of result

This paper mainly focuses on finding the accuracy of decision tree and C 5.0 algorithm and the yield of the crop from the available sample soil dataset collected from the Krishnagiri district in Tamil Nadu. 249 tuples of soil dataset with 24 attributes is used and the results are listed below:

1. Density of CA - calculated to find the accuracy soil data set using decision tree C5.0 Algorithm.
2. Histogram of CA - to identify the range of values present from 249 soil dataset.
3. Pie chart of crop, pie chart of crop with percentage – to identify the high yield of the crop.

Fig 4: Pie chart of crops.

Fig 5: Pie chart of crops with percentage.

Table 3: Correlation of Soil dataset.

| CA       | MG       | SO4      | B       | PH       |
|----------|----------|----------|---------|----------|
| CA 257406.46269 48166.2031 20977.747960 43.00442091 118.70608239 |
| MG 48166.2031 79884.512178 83.726130 54.99498309 19.51296743 |
| SO4 20977.747960 83.726130 10243.872263 32.95642246 15.41082070 |
| B 43.00442091 54.99498309 32.95642246 1.42030760 0.04246016 |
| PH 118.706082 19.51296743 15.41082070 0.04246016 0.35253919 |
| Ec -22.067120 -7.750496 -1.564885 -0.178828 -0.03287019 |
| N -1417.377817 -1407.055172 216.327696 13.759522 -0.55447597 |
| P 12663.37602 4271.373494 1388.457666 3.93059522 7.83020793 |
| K -43858.10192 -9391.70532 -7249.100353 -34.52472794 -32.17214017 |

Above table describes about correlation of the soil data set, in this we calculated the correlation, for nutrients such as calcium, magnesium, sulphate, boron, potential hydrogen, electrical conductivity, nitrogen, phosphorus and potassium. Correlation is calculated to find the accuracy of the decision tree algorithm.

Fig 6: Boxplot diagram of crops.

Fig 7: Rplot diagram of the nutrients.

Fig 8: Scatter 3D plot diagram of nutrients.
Table 4: Accuracy of the algorithm.

| Pred    | Positive | Negative |
|---------|----------|----------|
| Positive| 20       | 180      |
| Negative| 10       | 1820     |

Accuracy: 0.9064  
95% CI: (0.8929, 0.9187)  
No Information Rate: 0.9852  
P-Value [Acc > NIR]: 1  
Kappa: 0.1521  
Mcnemar's Test P-Value: <2e-16  
Sensitivity: 0.666667  
Specificity: 0.910000  
Pos Pred Value: 0.100000  
Neg Pred Value: 0.994536  
Prevalence: 0.014778  
Detection Rate: 0.009852  
Detection Prevalence: 0.098522  
Balanced Accuracy: 0.788333

‘Positive’ Class: positive

4. Calculated the mean, median, variance, correlation and quartiles – to find the accuracy of the soil dataset.
5. Boxplot diagram of crop – using R – Tool.
6. Scatter Diagram of CA.
7. 3D scatter diagram of available nutrients (N,P,K) attributes.
8. Rplot Diagram of the Nutrients.

The estimated results using Decision Tree Algorithm and C 5.0 Algorithmic technique are given in the Table 3. This work can be extended to all the attribute of the nutrient in the soil dataset in future.

CONCLUSION

Data mining is one of the emerging fields of technology nowadays used in agricultural field. Data mining algorithm techniques are used in the prediction of crop. This paper focuses on finding the accuracy of Decision Tree Algorithm and C5.0 Algorithm using sample soil data set collected from district agricultural office at Krishnagiri district in Tamil Nadu, the sample soil data set taken it from the dimension of 249 tuples and 24 attributes and the results of Decision Tree Algorithm Balanced Accuracy is 0.788333 and the Crop Yield prediction are Ragi-48%, Paddy-28% and Thuvvari-24%. From the experiment and results the accuracy of Decision Tree Algorithm is .78 and high yield predicted Crop is Ragi. This research uses R-tool to implement data mining algorithms on soil dataset. This paper focused only on three crops to find the crop that produce high yield, from sample soil dataset collected from Krishnagiri district, but in future we planned to take at least 6 to 8 crops to find the high yield of the Crop in Krishnagiri district. Availability of these nutrients and micronutrients can be used to decide its effects on the yielding capability of crops. Classification algorithms like J48 and ID3 gives the better prediction result. In future this work can be extended to find the accuracy of j48 algorithm and feature selection algorithms- applied predictive model is used to predict the crop yield and accuracy and then by combining these two j48 + C 5.0 algorithm to find the better accuracy of soil data set based on the prediction of crop yield.

REFERENCES

Hemageetha, N., Nasira, G.M. (2016). Analysis of the soil data using classification techniques for agricultural purpose. International Journal of Computer Sciences and Engineering. 4(6).

Hari Ganesh, S., S Jayasudha. (2015). Data mining technique to predict the accuracy of the soil fertility. International Journal of Computer Science and Mobile Computing. 7: 330-334.

Kanjana Devi, P., Shenbagavadivu, S. (2016). Enhanced crop yield prediction and soil data analysis using data mining. International Journal of Modern Computer Science. 4(6).

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).

Manjula, E., Djodilatouchamy, S. (2017). A model for prediction of crop yield. International Journal of Computational Intelligence and Informatics. 6(4): 298-305.

Rajeswari, V., Arun, K. (2016). Analysing soil data using data mining classification techniques. Indian Journal of Science and Technology, 9(19).

Samundeeswari, K., Srinivasan, K. (2018). Crop yield prediction and soil data analysis using data mining techniques in Krishnagiri district. International Journal of Computer Science and Engineering. 6(8): 49-55.

Samundeeswari, K., Srinivasan, K. (2017). Data mining techniques in agriculture, Prediction of Soil Fertility. International Journal of Scientifc and Engineering Research. 8(4): 45-51.

Snehal, S.D. (2019). Agricultural crop yield prediction using artificial neural network approach. International Journal of Innovative Research in Electrical, Electronic. 2(1): 683-686. https://krishnagiri.nic.in/about-district/history.