Understanding Decision Tree Algorithm by using R Programming Language

Akshat Sharma[#1], Anuj Srivastava[#2]
[#]*Computer Science Department, Integral University*
Lucknow, Uttar Pradesh, India

***Abstract*— Decision Tree is one of the most efficient technique to carry out data mining, which can be easily implemented by using R, a powerful statistical tool which is used by more than 2 million statisticians and data scientists worldwide. Decision trees can be used in a variety of disciplines, such as for predicting which patient characteristics are associated with high risk of a disease. When used together, we can find the relevant set of data, from the large data stored in the Enterprise Data Warehouses (EDWs). This provides new opportunities for organizations to derive new value from their most valuable and abundantly available asset: information, to create a competitive edge.**

***Keywords*— Decision Tree Algorithm, R Programming Language, Data Mining.**

I. INTRODUCTION

The first three phases of Data Analytics Lifecycle- discovery, data preparation, and model planning, involve various aspects of data exploration. The success of a data analysis project requires a deep understanding of the data, it requires a tool for data mining and presenting the data. We can use decision tree as a tool for data mining and R for presenting the data.

Data Mining is an analytic process designed to explore data (usually large amounts of data - also known as "big data") in search of consistent patterns and/or systematic relationships between variables, and then to validate the findings by applying the detected patterns to new subsets of data. The ultimate goal of data mining is prediction. The most common type of data mining is Predictive data mining and it is the one that has the most direct business applications. [1]

Fundamental learning methods that appear in applications related to data mining are-

1. Clustering
2. Association Rules
3. Regression
4. Classification
5. Time Series Analysis
6. Text Analysis

A. Clustering

Clustering is the use of unsupervised techniques for grouping similar objects. No predictions are made in clustering methods, instead the similarities between objects according to their object attributes are found and the similar objects are grouped into clusters. K-means is a popular clustering method.

B. Association Rules

It is an unsupervised learning method. Association rules is descriptive, not predictive method, used for discovering interesting relationships hidden in large dataset.

C. Regression

In simple terms, regression analysis is an explanatory tool that can be used to identify the input variables that have the greatest statistical influence on the outcome. There are two regression techniques- linear regression, and logical regression.

D. Classification

Classifiers are used in classification learning, hence the name. The primary task done by classifiers is to assign class labels to new observations. This technique is mostly used for prediction purposes. Two fundamental classification learning methods are decision trees and Naïve Bayes.

E. Time Series Analysis

In time series analysis method we attempt to model the underlying structure of observations taken over time, for this we use a time series. A time series is an ordered sequence of equally spaced values over time.

F. Text Analysis

Text analysis method constitutes of the representation, processing, and modelling of textual data to derive useful insight.

II. R LANGUAGE

R is a programming language and software framework for

statistical analysis and graphics. R is available for use under the GNU General Public License. [2]

R was created by Ross Ihaka and Robert Gentleman at the University of Auckland, New Zealand, and is currently developed by the R Development Core Team. R is an open source, portable, polymorphic, mainly command driven and a very powerful statistical programming language.

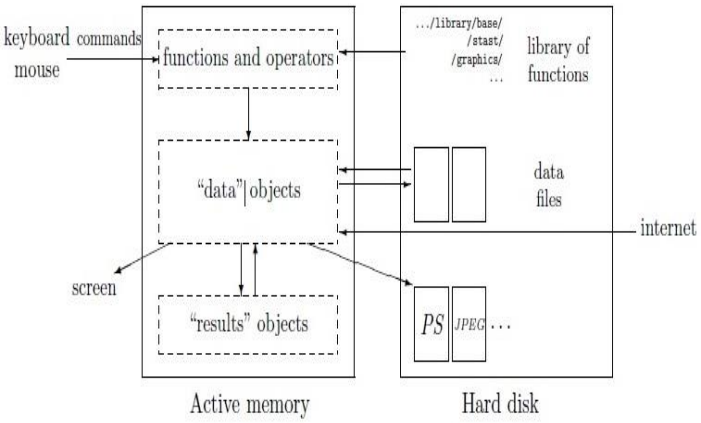

Figure 1. A schematic view of how R works. [3]

Figure1 aptly shows the working of R. In R, all actions are done on objects stored in the active memory of the computer. Temporary files are not used in R. To input and output data and results, the reading and writing of files are used. Data files can be read from the local disk or from a remote server through internet.

Programming with Big Data in R (pbdR) is a series of R packages and an environment for statistical computing with big data by using high performance statistical computation.

R software uses a command- line interface (CLI), which is similar to the BASH shell in Linux. To improve the ease of writing, executing, and debugging R code, some popularly used GUIs include the RStudio [4], the R commander [5], and Rattle [6].

The RStudio will be used as the GUI to build the R code for decision tree. The RStudio consists of four main window panes- Scripts, Workspace, Plots, and Console. Scripts serves as an area to write and save R code, Workspace lists the datasets and variables used in the R environment, Plots window is used to display the plots generated by the code and provides a mechanism to export the plots, Console shows the history of the executed RCode and the output.

The R console is the most important tool for using R. The commands that you type in the console are called expressions. The R system use an interpreter, it reads the expressions and respond with a result or an error message. By default, R will display a greater than sign (">") in the console when R is waiting for you to enter a command into the console, this is called a prompt.

R is packaged with a small number of essential packages by default, and being open source, many R packages are contributed by users worldwide. Some packages are loaded by default with every session. The libraries shown in Table 1 are loaded on the R startup.

Package	Description
base	Base R functions
datasets	Base R datasets
grDevices	Graphics devices for base and grid graphics
graphics	R functions for base graphics
methods	Formally defined methods and classes for R objects
stats	R statistical functions
utils	R utility functions

Table 1. R packages, loaded on startup. [7]

III. DECISION TREE

A decision tree is also called a prediction tree. A decision tree uses a structure to specify sequences of decisions and consequences. Given input X={X1, X2,..,Xn}, the goal is to predict a response or output variable Y. [8] Each member of the set {X1,X2,…,Xn} is called an input variable.

The prediction can be achieved by constructing a decision tree with test points and branches. A decision is made at each test point, to pick a specific branch and traverse down the tree

Decision trees can be used in a variety of disciplines, such as: On the basis of individual characteristics deciding whether or not to offer a loan to an individual, predicting the rate of return of various investment strategies, predict whether or not send a direct mail to a potential customer, etc.

A decision tree consists of nodes, and thus form a rooted tree, this means that it is a directed tree with a node called root. There are no incoming edges on root node, all other nodes in a decision tree have exactly one incoming edge. An internal node is a node with an incoming edge and outgoing edges, internal node is also known as test node. Nodes with no outgoing edges are known as leaves or terminal nodes.

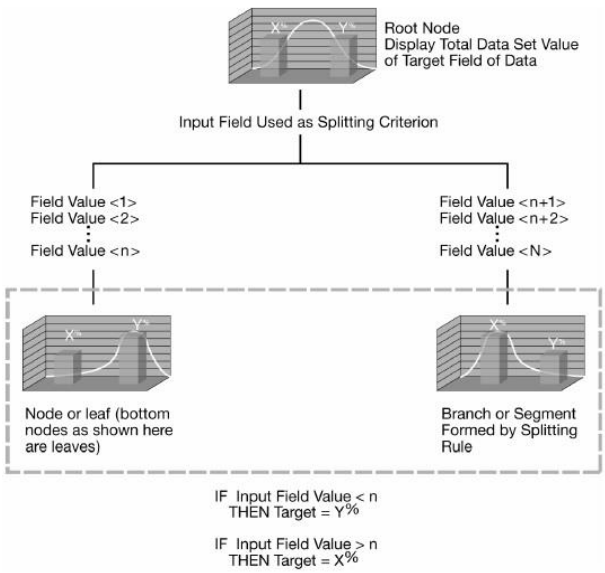

Figure 2. Illustration of the decision tree [9]

Decision trees are produced by algorithms that identify various ways of splitting a data into branch-like segments. The decision tree shown in Figure 2, clearly shows that decision tree can reflect both a continuous and categorical object of analysis.

Figure 3 describes a decision tree that reasons whether to send an alert or to send a warning when an error is triggered.

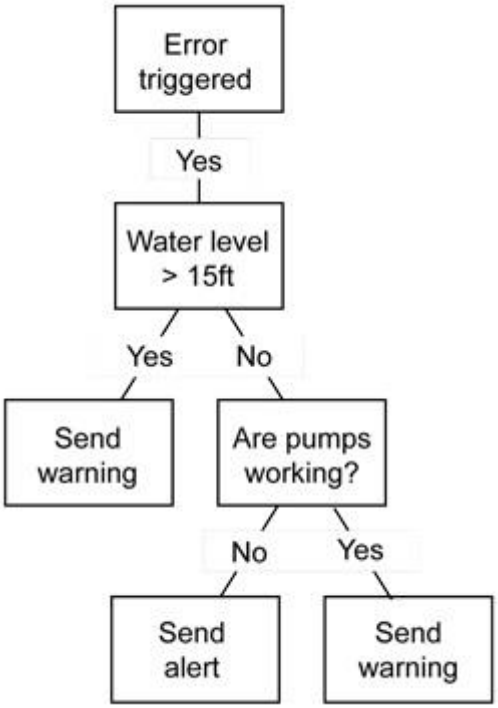

Figure 3. Decision tree presenting warning system.

In Figure 3, to depict the tree in more clear way, we can represent the internal nodes as circles, and leaves as triangles, to tell them apart in one glance. Each node is labeled with the attribute it tests, and its branches are labeled with its corresponding values. The leaf nodes represent class labels, i.e. the outcome of all the prior decisions.

A short tree can be created, by limiting the number of splits. Short trees are often used as components in ensemble methods, such as, random forest, bagging, and boosting. The simplest short tree is called a decision stump. Decision stump is a decision tree with root node immediately connected to the leaf nodes. A decision stump makes a prediction based on the value of just a single input variable.

IV. The General Decision Tree Algorithm

The objective of a decision tree algorithm is to construct a tree 'T' from a training set 'S'. If all the records in 'S' belong to some class 'C', or if 'S' is sufficiently pure, then that node is considered a leaf node and assigned the label 'C'. The purity of a node is defined as its probability of the corresponding class. [10]

The algorithm constructs subtrees for the subsets of S recursively until one of the following criteria is met: [11]

1. All the leaf nodes in the tree satisfy the minimum purity threshold.
2. The tree cannot be further split with the preset minimum purity threshold.
3. Any other stopping criterion is satisfied (such as the maximum depth of the tree).

Entropy and information gain are common technical terms/notations used in decision trees. While constructing a decision tree we first choose the most informative attribute. Entropy-based methods can be used to choose the most informative attribute. The entropy methods select the most informative attribute based on, entropy and information gain.

Entropy is a measure of the number of random ways in which a system may be arranged. For a data set 'S' containing 'n' records the information entropy is defined as,[12]

$$Entropy(S) = -\sum P_1 \log_2 P_1$$

(Here 'P_1'is the proportion of 'S' belonging to class 'I').

Gain or the information gain is the change in information entropy from a prior state to a state that takes some information. The information gain of example set 'S' on attribute 'A' is defines as,[13]

$$Gain(S,A) = Entropy(S) - \sum_{v \in Value(A)} \frac{|Sv|}{|S|} Entropy(S_v)$$

Where 'Value(A)' is the set of all possible values of attribute 'A', 'S_v' is the subset of' 'S' for which attribute 'A' has the value 'v', '$|S_v|$' is the number of elements in 'S_v' and |S| is the number of elements in 'S'.

The algorithm for decision tree growth phase is as

under:[14]

BuildTree (data set S)
If all records in S belong to the same class,
return;
for each attribute A_i
evaluate splits on attribute A_i;
use best split found to partition S into S_1 and S_2;
BuildTree(S_1);
BuildTree(S_2);
endBuildTree;

Multiple algorithms exist to implement decision trees, some popular algorithms include ID_3, $C_{4.5}$, and CART.

ID_3 (or Iterative Dichotomiser$_3$) [15] was developed by John Ross Quinlan. It is one of the first decision tree algorithms.

$C_{4.5}$ algorithm [16] is an improvement over the original ID3 algorithm. It can handle missing data. The overfitting problem in ID_3 is addressed by the $C_{4.5}$ algorithm.

CART (or Classification and Regression Trees) [17] is similar to $C_{4.5}$ and can handle continuous attributes, CART takes use of the Gini diversity index.

Gini index for a data set 'S' is defined as, [18]

$gini(S) = 1 - \sum P_I^2$.

V. Decision Trees in R

In R, 'rpart' is for modelling decision trees, and an optional package 'rpart.plot' enables the plotting of a tree. [19] The rpart package can be used for classification by decision trees and can also be used to generate regression trees.

To grow a tree, use [20]

rpart (formula, data=, method=, control=)

Where,

formula	is in the format: outcome ~ predictor1+predictor2+ect.
data=	specifies the dataframe
method	"class" for a classification tree, "anova" for a regression tree
control=	optional parameters for controlling tree growth.

Table 2. Terms used in command to grow a tree. [20]

With the help of following functions we can examine the results,

printcp(*fit*)	display cp table
plotcp(*fit*)	plot cross-validation results
rsq.rpart(*fit*)	plot approximate R-squared and relative error for different splits.
print(*fit*)	print results
summary(*fit*)	detailed results including surrogate splits
plot(*fit*)	plot decision tree
text(*fit*)	label the decision tree plot
post(*fit*,file=)	create postscript plot of decision tree

Table 3. Functions to examine the results. [21]

Following is a simple example of decision trees in R, using rpart package: [22]

```
1   library(rpart)
2   raw.orig < - read.csv(file="c:\\rsei212_chemical.txt", header=T, sep="\t")
3
4   # Keep the dataset small and tidy
5   # The Dataverse: hdl:1902.1/21235
6   raw = subset(raw.orig, select=c("Metal","OTW","AirDecay","Koc"))
7
8   row.names(raw) = raw.orig$CASNumber
9   raw = na.omit(raw);
10
11  frmla = Metal ~ OTW + AirDecay + Koc
12
13  # Metal: Core Metal (CM); Metal (M); Non-Metal (NM); Core Non-Metal (CNM
14
15  fit = rpart(frmla, method="class", data=raw)
16
17  printcp(fit)  # display the results
18  plotcp(fit)  # visualize cross-validation results
19  summary(fit)  # detailed summary of splits
20
21  # plot tree
22  plot(fit, uniform=TRUE, main="Classification Tree for Chemicals")
23  text(fit, use.n=TRUE, all=TRUE, cex=.8)
24
25  # tabulate some of the data
26  table(subset(raw, Koc>=190.5)$Metal)
```

Simple R code of Classification Tree for Chemicals. [22]

The output of the above code will be,

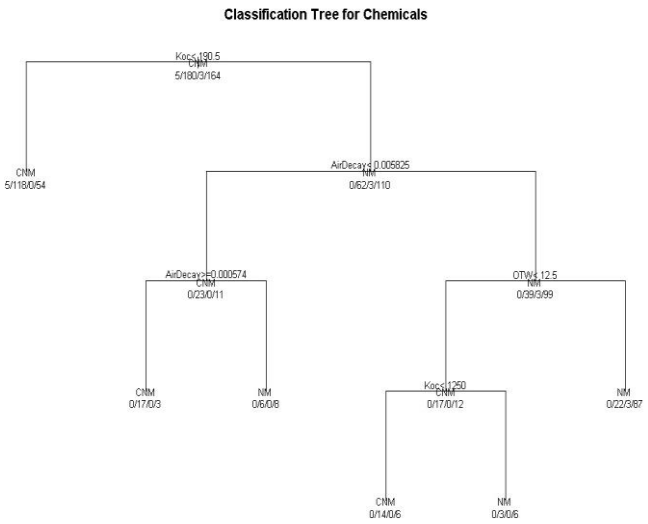

Figure 4. Output of R code for Classification Tree for Chemicals [22].

VI. Conclusion

Decision trees, like all other classifiers, have its own set of advantages and disadvantages. Table 4 offers a list of conditions when use of decision trees pose an advantage.

Concerns	Recommended Method(s)
Output of the classification should include class probabilities in addition to the class labels.	Decision tree, logistic regression.
Analysts want to gain an insight into how the variables affect the model.	Decision tree, logistic regression.
Some of the input variables might be correlated.	Decision tree, logistic regression.
Some of the input variables might be irrelevant.	Decision tree, naïve Bayes.
The data contains categorical variables with a large number of levels.	Decision tree, naïve Bayes.
The data contains mixed variable types.	Decision tree, logistic regression.
There is nonlinear data or discontinuities in the input variables that would affect the output.	Decision tree.

Table 4. Conditions for Decision Tree used as suitable Classifier. [23]

The decision tree cannot be used as classifier when, the problem is high dimensional. In that case, the problem can be classified by using Naïve Bayes.

The use or R for implementing decision trees for classification in data mining is highly over SAS and Matlab because, R is open source, has a large library support and support visualization, while being inexpensive in comparison to both. The disadvantage of using R is that it has a steep learning curve.

References

[1] Data Mining Techniques. [Online]. Available: http://documents.software.dell.com/Statistics/Textbook/Data-Mining-Techniques [Accessed 21 November 2015].

[2] EMC Education Services, "Data Science and Big Data Analytics," in *Review of Basic Data Analytic Methods Using R*, January 2015, pg. 64.

[3] Emmanuel Paradis, "R for Beginners", CRAN. Figure 1, pp. 8. [Online]. Available: https://cran.r-project.org/doc/contrib/Paradis-rdebuts_en.pdf [Accessed 22 August 2015].

[4] RStudio. [Online]. Available: https://www.rstudio.com/products/rstudio/ [Accessed 21 November 2015].

[5] J. Fox and M. Bouchet-Valat, "The R Commander: A Basic-Statistics GUI for R", CRAN. [Online]. Available: http://socserv.mcmaster.ca/jfox/Misc/Rcmdr/ [Accessed 21 November 2015].

[6] G. Williams, M.V. Culp, E. Cox, A. Nolan, D. White, D. Medri, and A. Waljee, "Rattle: Graphical User Interface for Data Mining in R," CRAN. [Online]. Available: https://cran.r-project.org/web/packages/rattle/index.html [Accessed 21 November 2015].

[7] Adedin Culhane, Harvard School of Public Health, BIO503, January 2013, "Introduction to Programming and Statistical Modelling in R". [Online]. Available: http://isites.harvard.edu/fs/docs/icb.topic1202070.files/Bio503_winter13.pdf [Accessed 22 August 2015].

[8] EMC Education Services, "Data Science and Big Data Analytics," in *Advanced Analytical Theory and Methods: Classification*, January 2015, pp. 192.

[9] "Decision Trees- What Are They?", figure1.1, pg. 3. [Online]. Available: http://support.sas.com/publishing/pubcat/chaps/57587.pdf [Accessed on 26/11/15].

[10] EMC Education Services, "Data Science and Big Data Analytics," in *Advanced Analytical Theory and Methods: Classification*, January 2015, pg. 197.

[11] EMC Education Services, "Data Science and Big Data Analytics," in *Advanced Analytical Theory and Methods: Classification*, January 2015, pg. 197.

[12] Venkatadri.M, Lokanatha C. Reddy. (2010, Apr.-2010, Sept.). A comparative study on decision tree classification algorithms in data mining. *International Journal of Computer Application in Engineering, Technology and Sciences (IJ-CA-ETS).* [Online]. *1.1*, pg. 1. Available: https://www.academia.edu/

[13] Venkatadri.M, Lokanatha C. Reddy. (2010, Apr.-2010, Sept.). A comparative study on decision tree classification algorithms in data mining. *International Journal of Computer Application in Engineering, Technology and Sciences (IJ-CA-ETS).* [Online]. *1.1*, pg. 1. Available: https://www.academia.edu/

[14] Venkatadri.M, Lokanatha C. Reddy. (2010, Apr.-2010, Sept.). A comparative study on decision tree classification algorithms in data mining. *International Journal of Computer Application in Engineering, Technology and Sciences (IJ-CA-ETS).* [Online]. *Fig. 1*, pg. 2. Available: https://www.academia.edu/

[15] J. R. Quinlan, "Induction od Decision Trees," *Machine Learning*, vol. 1, no. 1, pp. 81-106, 1986.

[16] J. R. Quinlan, $C_{4.5}$: *Programs for Machine Learning*, Morgan Kaufmann, 1993.

[17] L. Breiman, J. H. Friedman, R. A. Olshen, and C. J. Stone, *Classification and Regression Trees*, Belmont, CA: Wadsworth International Group, 1984.

[18] Venkatadri.M, Lokanatha C. Reddy. (2010, Apr.-2010, Sept.). A comparative study on decision tree classification algorithms in data mining. *International Journal of Computer Application in Engineering, Technology and Sciences (IJ-CA-ETS).* [Online]. *1.1*, pg. 1. Available: https://www.academia.edu/

[19] EMC Education Services, "Data Science and Big Data Analytics," in *Advanced Analytical Theory and Methods: Classification*, January 2015, pg. 206.
[20] Data Mining Algorithms In R/Classification/Decision Trees. [Online]. Available: https://en.wikibooks.org/wiki/Data_Mining_Algorithms_In_R/Classification/Decision_Trees [Accessed 27 November 2015].
[21] Tree-Based Models. [Online]. Available: http://www.statmethods.net/advstats/cart.html [Accessed 27 November 2015].
[22] A Brief Tour of the Trees and Forests. [Online]. Available: http://www.r-bloggers.com/a-brief-tour-of-the-trees-and-forests/ [Accessed 27 November 2015].
[23] EMC Education Services, "Data Science and Big Data Analytics," in *Advanced Analytical Theory and Methods: Classification*, January 2015, pg. 229.