Chapter 7
Implementation of Data Mining Algorithm With R

C. Deisy
Thiagarajar College of Engineering, India

Mercelin Francis
Thiagarajar College of Engineering, India

ABSTRACT

R is a programming language that uses command-line scripting for graphical and statistical analysis and representation and finally generating a report. It is a free, open source, powerful, and highly extensible tool for data analysis. It consists of a large repository of intermediate tools for statistical and graphical analysis of data which utilizes conditional loops and user-defined functions with input and output capabilities. Statistical and analytical techniques are developed with R for various decision-making processes like forecasting, social media analytics, text mining, and so on. The chapter focuses on the basics of R, data storage elements, and its manipulation. It also highlights the usage of the machine learning algorithms for prediction, clustering, and classification. Applications like text mining are implemented to extract various patterns or rules based on the scenario. Illustrations are explained providing a base for developing many applications applying the basic concepts of R.

INTRODUCTION

This chapter deals with the Implementation of Data mining algorithm with R Program. Data mining is the task to extract the interesting rules or patterns or regularities or constraints from large data which is previously unknown. The advancement in information technology and the social media leads to a very big challenge for handling the diversified data. The process of analyzing (mining) the huge data (text, audio, video, image and graph data)and to retrieve (extract or excavate) the knowledge or information from the data is the major role of data mining. Based on the type of data we can categorize data mining into Web mining, text mining, image mining, and content mining. The scanning and mining of text, pictures and graphs of a Web page to determine the relevance of the content to the search query is called content mining.

DOI: 10.4018/978-1-5225-4999-4.ch007

The application of data mining is to identify the customer behavior understanding for retail shop. Also helps to find the fraud detection and stock trading in real time decision making systems. It helps to identify the inventory management, and pricing of a product in business decision making systems. The amazon.com system use recommendation algorithms to personalize the online store for each customer. To identify the customer's behaviors it uses customer's interest as input to generate a list of recommended items (Applications of Data Mining, n.d.).

R is a free open source software package, available under the GNU General Public License. Obtaining R, its installation and building R on the system is the preliminary steps performed before executing the basic commands (Gardener, 2015; Peng, 2015; Martin, 2009). R was created by Ross Ihaka and Robert Gentleman, and currently developed by R development core team. It runs on Microsoft, UNIX and Macintosh. It is increasingly applied in business analytics, to visualize the data in excellent graphical output. R programming language is used by data analyst and others who want to make statistical analysis of data and infer insights from data using mechanisms, such as regression, clustering, classification, and text analysis (Tutorials point, 2016).

R provides a wide variety of statistical, machine learning (linear and nonlinear modeling, classic statistical tests, time-series analysis, classification, clustering) and graphical techniques, and is highly extensible. R has various built-in as well as extended functions for statistical, machine learning, and visualization tasks such as: Data extraction, Data cleaning, Data loading, Data transformation, Statistical analysis, Predictive modeling and Data visualization. It is a cross platform with a very wide, large and ever-growing user community support who adds new packages every day.

The first topic in this chapter handles the basic R programming: The basic mathematical operations using R, relational Operations and logical operations (Sarkar, 2012; Knell, 2014; Martin, 2009).

Second topic covers the creation and indexing of Numeric vectors and Matrices using. It also deals with importing and exporting data, Text variable and Vectors, Logical variable and vectors, Matrices and vectorised calculations using Apply, Subsetting Vectors, Subscripts and Matrices, Managing Workspace, Data Frames, Lists, Saving a workspace file using R, Importing and Exporting Data, Installing, loading and unloading packages R (R Development Core Team, 2005; Gardener, 2015; Martin, 2009).

Third topic covers plotting of various graphs using the R programming. It also handles with descriptive statistical analysis for pie chart, Bar chart and Histogram (R Development Core Team, 2005; Gardener, 2015). Also identifies the relation between categorical and numerical variables, how to calculate the covariance and correlations.

Fourth topic handles on R Statistical operations, such as mean, min, max. Machine learning operations, such as linear regression is a classic technique to identify the scalar relationship between two or more variables by fitting the state line on the variable values. That relationship will help to predict the variable value for future events. Sales forecasting of products or services and predicting the price of stocks can be achieved through this regression. R provides this regression feature via the lm method, which is by default present in R (Gardener, 2015; Kohl,2015).

The next topic deals with the Classification and cluster Analytics using R applied on text mining. Both are machine-learning techniques. Classification is used for labeling the set of observations provided for training examples. It can classify the observations into one or more labels. The likelihood of sales, online fraud detection, and cancer classification (for medical science) are common applications of classification problems. Google Mail uses this technique to classify e-mails as spam or not. Clustering technique is all about organizing similar items into groups from the given collection of items. User segmentation and image compression are the most common applications of clustering. Market segmentation, social