IJournals: International Journal of Software & Hardware Research in Engineering
ISSN-2347-4890
Volume 3 Issue 9 September, 2015

Data Mining Using R

Author: Dhwani Raval

Affiliation: VIT University

E-mail: dhwani20raval@gmail.com

ABSTRACT

Currently large organizations are affected by sudden growth in data and common software tools are used for managing, capturing and processing on dataset. Today main challenge is that analyzes all the datasets and gets more accurate and useful information in short period of time. This paper gives an insight on data mining using R statistical software. This paper proposes three stages for data mining process 1) the initial exploration, 2) model building or pattern identification with validation or verification and, 3) deployment means prediction of application

Keywords: Data Mining, R, Big Data

1. INTRODUCTION

Generally, data mining (sometimes called data or knowledge discovery) is the process of analyzing data from different perspectives and summarizing it into useful information - information that can be used to increase revenue, cuts costs, or both. Data mining software is one of a number of analytical tools for analyzing data. It allows users to analyze data from many different dimensions or angles, categorize it, and summarize the relationships identified. Technically, data mining is the process of finding correlations or patterns among dozens of fields in large relational databases [1].

Data Mining is an analytical process designed to explore data, usually large amount of data that typically related to business or market, and this type of data also called "Big Data". In search of consistent patterns or systematic relationships between variables, and then to validate the finding by applying the detected patterns to new subsets of data.

2. COMPONENTS OF DATA MINING

Data Mining consists of five major components.

1. Extract, transform, and load transaction data onto the data warehouse system.
2. Store and manage the data in a multidimensional database system.
3. Provide data access to business analysts and information technology professionals.
4. Analyze the data by application software.
5. Present the data in a useful format, such as a graph or table.

3. ANALYSIS IN DATA MINING

There are various types of analysis tool available in Data Mining.

3.1 Artificial neural networks

Non-linear predictive models that learn through training and resemble biological neural networks in structure.

3.2 Genetic algorithms

Optimization techniques that use process such as genetic combination, mutation, and natural selection in a design based on the concepts of natural evolution.

3.3 Decision trees

Tree-shaped structures that represent sets of decisions. These decisions generate rules for the classification of a dataset. Specific decision tree methods include Classification and Regression Trees (CART) and Chi Square Automatic Interaction Detection (CHAID). CART and CHAID are decision tree techniques used for classification of a dataset. They provide a set of rules that you can apply to a new (unclassified) dataset to predict which records will have a given outcome. CART segments a dataset by creating 2-way splits while CHAID segments using chi square tests to create multi-way splits. CART typically requires less data preparation than CHAID.

ISSN-2347-4890

Volume 3 Issue 9 September, 2015

3.4 Nearest neighbour method

A technique that classifies each record in a dataset based on a combination of the classes of the k record(s) most similar to it in a historical dataset (where k > 1). Sometimes called the k-nearest neighbour technique.

3.5 Rule induction

The extraction of useful if-then rules from data based on statistical significance.

3.6 Data visualization

The visual interpretation of complex relationships in multidimensional data. Graphics tools are used to illustrate data relationships.

Data Mining is applicable in various area as we can see from Fig. 1.

4. PROPOSED WORK

The proposed work divides in 3 stages; all rest part is explained in details all proposed methods. From the above I mention different level of analysis, and data visualization is one of them. So for data visualization we use R software as well we used for different Data Mining technique also.

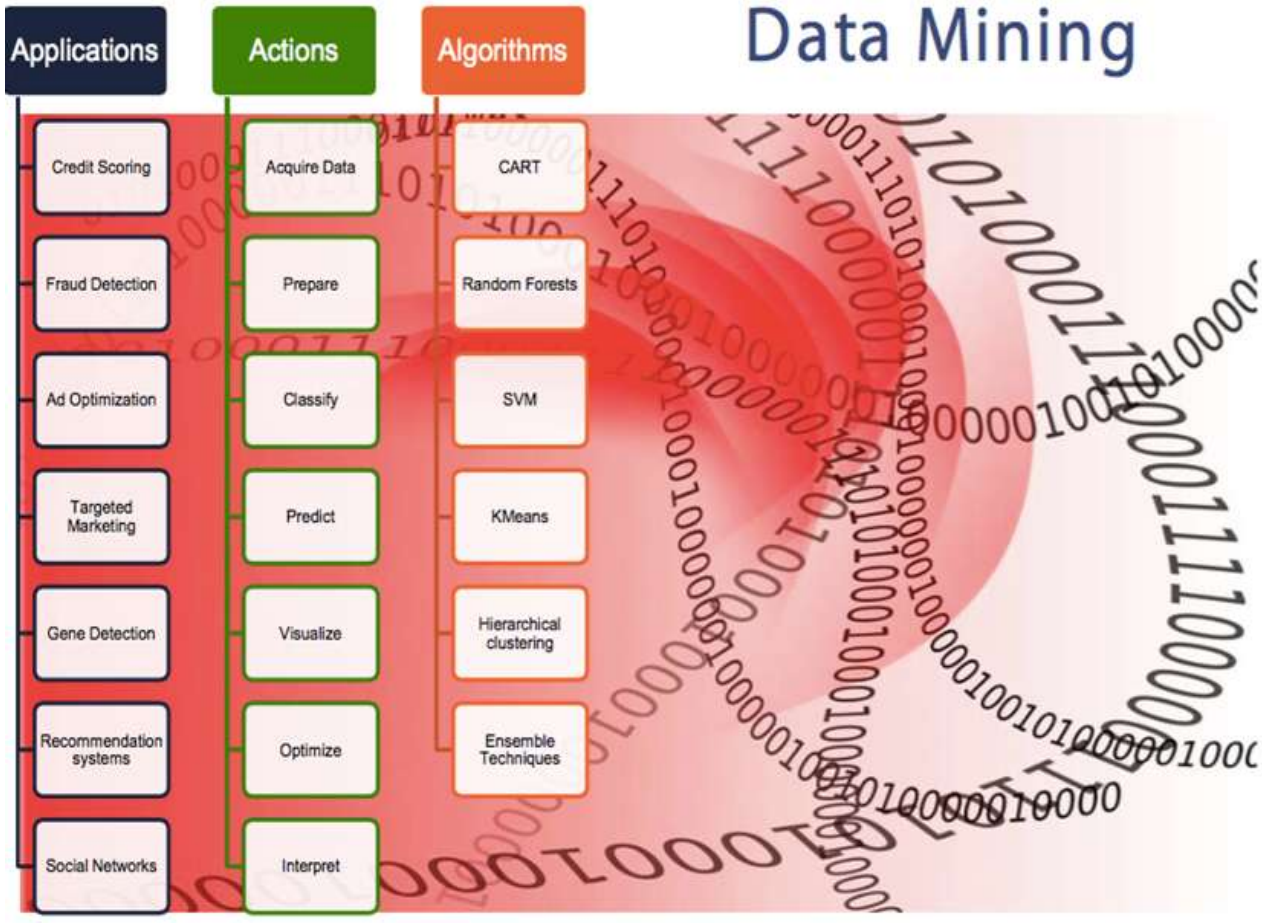

Fig 1: Data Mining applications, actions and algorithms

- Step 1 (Exploration): In this step by default start with getting data from any website or also called data preparation and in this include data cleaning, data transformation. If dataset has many parameter or field than using R tool we can also modify and used only important field so able to remove un-necessary field. Depending on nature of analytics

- Step 2 (Validating and Model developing): In this stage various model and different type of validation techniques are included. This is a look like very simple stage but sometime it is very difficult and elaborate process. This technique mainly used for predictive data mining – for example Voting Analysis, Stock market analysis etc.

- Step 3 (Deploying): This is a final stage of the Data Mining process. Data Mining has become very popular as a business information management tool where we use R Tool using R Language.

This stage is very important for developing predictions from previous stage and makes it from new data. Recently, there has been increased interest in developing new analytic techniques specifically designed to address the issues relevant to business Data Mining, but Data Mining is still based on the conceptual principles of statistics including the traditional Exploratory Data Analysis (EDA) and modeling and it shares with them both some components of its general approaches and specific techniques.

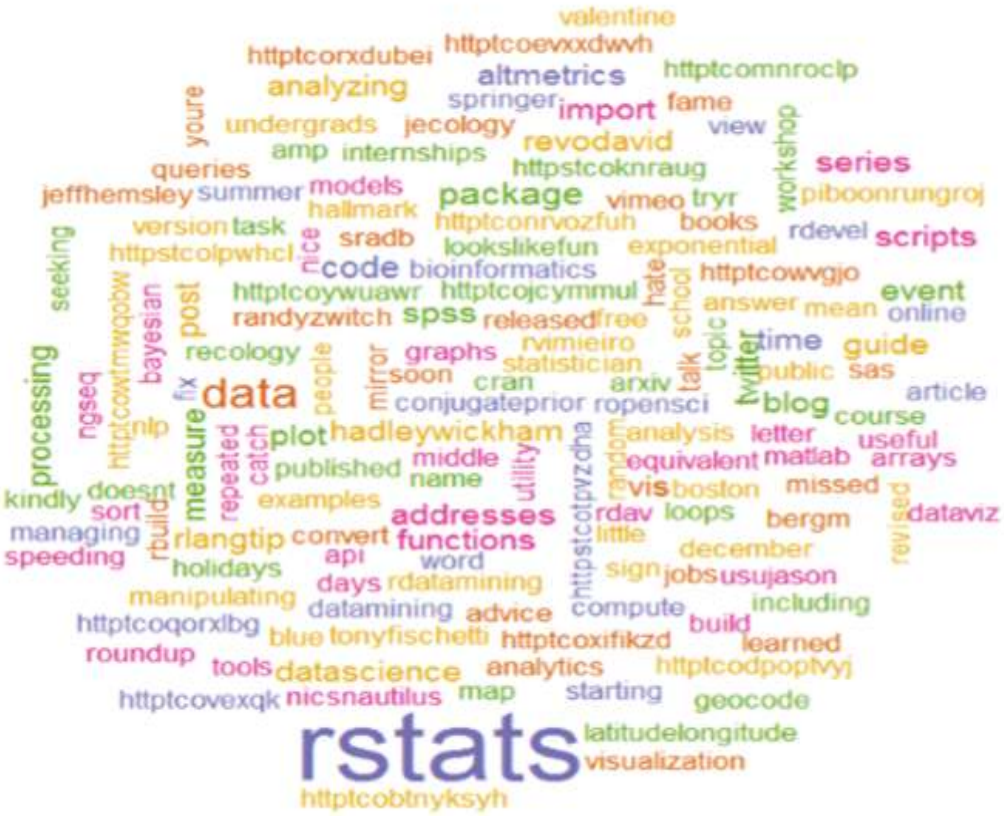

Fig 2: R Package used for Data Mining and Visualization

5. CONCLUSION

This paper proposes Data mining processes for big data using R statistical software. The proposed method is composed by total three stages. In very first stage, datasets are extracted from different sites. In second stage, different models are developed and validated as per requirements. In third stage developing task are carried out according to requirement and visual analysis are done using R statistical software. The results of propose method is providing useful for organization and medical (Healthcare) field.

6. REFERENCES

[1]. Y. Zhao, R and Data Mining, 1st ed., Academic Press, 2012.

[2]. Y. Zhao, J. Cen, Data Mining Application with R, Elsevier, 2012.

[3]. V. Kumar, M Joshi, E Han, P Tan and M Steinback, High Performance Computing for Computational Science, 2002, Portugal.

[4]. G. Simon, H Xiong, E Eilertson, and V Kumar, Scan Detection – A Data Mining Approach 2006,

[5]. G. Williams, Rattle: A Data Mining GUI for R, The R Journal, Vol. 1/2, Dec 2009, pg 45-55.

[6]. J. Hsu, "Data Mining Trends and Developments: The Key Data Mining Technologies and Applications for the 21st Century", in The Proceedings of the 19th Annual Conference for Information Systems Educators.

[7]. M. Venkatadari, L. Reddy, "A Review on Data Mining From Past to Future", International Journal of Computer Applications, 2011, vol. 15, No. 7, pp. 19-22.

