INTERNATIONAL JOURNAL OF RESEARCHES IN BIOSCIENCES, AGRICULTURE AND TECHNOLOGY
© VISHWASHANTI MULTIPURPOSE SOCIETY (Global Peace Multipurpose Society) R. No. MH-659/13(N)
www.vmsindia.org

Big Data Analytics Using R

Madhura A.Chinchmalatpure[1] and Mahendra P. Dhore[2]
[1]Department of Computer Science, SSESA, Science College, Congress Nagar, Nagpur, (MS)-India,
madhura.naralkar@gmail.com
2Department of Electronics & Computer Science, RTM Nagpur University, Nagpur, (MS)-India,
mpdhore@rediffmail.com

Abstract
R is an open source programming language and software environment for statistical computing and graphics that is supported by the R Foundation for Statistical Computing. The R language is widely used among statisticians and data miners for developing statistical software and data analysis. The process of converting data into knowledge, insight and understanding is Data analysis, which is a critical part of statistics. For the effective processing and analysis of big data, it allows users to conduct a number of tasks that are essential. R consists of numerous ready-to-use statistical modelling algorithms and machine learning which allow users to create reproducible research and develop data products.

Keyword- *R, Big Data Analytics, Electronic Health Record(EHR)*

1. INTRODUCTION

I) Big Data Analytics:

Big data is a term defining collection of large datasets. After interaction with patients here we implement electronic health record(EHR) which is highly capable of storing voluminous data in database and this includes patient's previous medical data, laboratory test values, current treatment given to patient, doctors prescription, diagnostic reports, pharmacy information health insurance related data, medical journals are used to proper investigate and analysis.

II) Purpose and Categories:

Data analytics addresses information obtained through comment, measurement, or tests about a phenomenon of interest. The following lists only a few potential purposes:

1) To generalize and deduce the data and determine how to use it.

2) To check whether the data are genuine.

3) To give guidance and contribution in decision making system.

4) To identify and conclude reasons for fault.

5) To forecast what will occur in the future.

Descriptive Analytics: exploits historical data to describe what occurred in past. For instance, a regression technique may be used to find simple trends in the datasets, visualization presents data in a meaningful fashion, and data modeling is used to collect, store and cut the data in an efficient way. Descriptive analytics is typically associated with business intelligence or visibility systems .

Predictive Analytics: focuses on predicting future probabilities and trends. For example, predictive modeling uses statistical techniques such as linear and logistic regression to understand trends and predict future out-comes, and data mining extracts patterns to provide insight and forecasts.

Prescriptive Analytics: addresses decision making and efficiency. For example, simulation is used to analyze complex systems to gain insight into system performance and identify issues and optimization techniques are used to find best solutions under given constraints.

III) R Programming Language:

R is a programming language and software environment for statistical analysis, data analytics ,scientific research, graphics representation, reporting. 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 freely available under the GNU General Public License, and pre-compiled binary versions are provided for various operating systems like Linux, Windows and Mac.

This programming language was named **R**, based on the first letter of first name of the two R authors (Robert Gentleman and Ross Ihaka), and partly a play on the name of the Bell Labs Language **S**.

R is an open-source implementation of S, and differs from S largely in its command-line. For statistical analyses, R has a broad set of facilities that has been specially constructed. As a result, R is said to be a very powerful statistical programming language. The open-source nature of R indicates that, as new techniques for statistics are developed, new packages for R usually become freely available very soon after. It consists of its own inbuilt statistical algorithms – the sheer amount of machine learning algorithms and mathematical models available to users in R and third-party packages is continues to grow.

R can also carry out important analyses that are difficult or next to impossible in many other such packages, including Generalized Additive Models, Linear Mixed Models and Non-Linear Models. R consists of broad range of graph-drawing tools, which makes it easy to produce standard graphs of your data. In traditional analysis, developing a statistical model takes more time than by performing the calculation by the computer. In case of Big Data this proportion is turned upside down. Big Data comes into picture when the CPU time for the calculation takes longer than the process of designing a model. Data sets that contain up to millions of records can easily be processed with standard R.

Data sets with almost one million to one billion records can also be processed in R, but requires some additional effort. Worldwide, millions of statisticians as well as data scientists use R in order to solve their most challenging problems in the field, right from quantitative marketing to computational biology.

R have become the most popular language for data science and an most essential tool for analytics-driven companies such as Google, Facebook, LinkedIn and Finance .

2. APPLICATION OF R PROGRAMMING LANGUAGE :

I) Data Science:

A data scientist is statistician with an extra asset,computer programming skills. Programming languages like R give a data scientist superpowers that allow them to collect data in realtime, perform statistical and predictive analysis, create visualizations and communicate actionable results to stakeholders.

II) Statistical computing:

R is the most popular programming language among statisticians. In fact, it was initially built by statisticians for statisticians. It has a rich package repository with more than 9100 packages with every statistical function you can imagine.

R's expressive syntax allows researchers - even those from non computer science backgrounds to quickly import, clean and analyze data from various data sources.

R also has charting capabilities, which means you can plot your data and create interesting visualizations from any dataset.

III) Machine Learning:

R has found a lot of use in predictive analytics and machine learning. It has various package for common ML tasks like linear and non-linear regression, decision trees, linear and non-linear classification and many more. Everyone from machine learning enthusiasts to researchers use R to implement machine learning algorithms in fields like finance, genetics research, retail, marketing and health care.

3. BENEFITS USING R

I) Package ecosystem.

One of R's strongest qualities is the vastness of package ecosystem .There's a lot of functionality that's built in and that's built for statisticians.

II) R is extensible

R provides rich functionality for developers to build their own tools and methods for analysing data. Lots of people aroused to it from other fields such as biosciences and even humanities. People can extend it without a need to ask permission.

III) Free software

At the time when R first came out, the biggest advantage of it was that it was free software. Every single thing and source code about it was available to look at.

IV) R's graphics and charting capabilities

For data manipulation and plotting the dplyr and ggplot2 packages, respectively have literally improved quality of life.

V) R's strong ties to academia

Any new research in the field probably has an associated R package to go. So R stays progressive. The caret package also offers a pretty smart way of doing machine learning in R

4. R's CHALLENGES

For all of its benefits, R has its share of shortcomings as follows-

I) Memory management

II) Speed

III) Efficiency.

These are probably the biggest challenges R faces. Also, people coming to R from other languages might also consider R odd. When working with very large data sets the design of the language can sometime lead to problems.

Data has to be stored in physical memory. But this can become a minor issue, as nowadays computers have plenty of memory. Abilities such as security were not built into the R language. Also, R cannot be embedded in a Web browser. You can't use it for Web-like or Internet-like apps. It was primarily next to impossible to use R as back-end server to perform calculations due to lack of security over the Web. For a long time, there was not a lot of interactivity in the language. Languages such as JavaScript still have to enter in to fill this gap. Although an analysis may be done in R, the furnishing of results might be accomplished in different language like JavaScript.

Beside data management operations and traditional statistic tasks – a wide range of data-mining algorithms like SVM, Neural Networks, Decision Trees etc.

5. Big Data Strategies in R

I) Sampling

If data is too big to be analyzed in complete, its' size can be reduced by sampling. Naturally, the question arises whether sampling decreases the performance of a model significantly. Much data is of course always better than little data. But according to Hadley Wickham's useR! talk, sample based model building is acceptable, at least if the size of data crosses the one billion record threshold.

If sampling can be avoided it is recommendable to use another Big Data strategy. But if for whatever reason sampling is necessary, it still can lead to satisfying models, especially if the sample is

- still (kind of) big in total numbers,
- not too small in proportion to the full data set,
- not biased.

II) Bigger hardware

R keeps all objects in memory. This can become a problem if the data gets large. One of the easiest ways to deal with Big Data in R is simply to increase the machine's memory. Today, R can address 8 TB of RAM if it runs on 64-bit machines. That is in many situations a sufficient improvement compared to about 2 GB addressable RAM on 32-bit machines.

III) Store objects on hard disc and analyze it chunkwise

As an alternative, there are packages available that avoid storing data in memory. Instead, objects are stored on hard disc and analyzed chunkwise. As a side effect, the chunking also leads naturally to parallelization, if the algorithms allow parallel analysis of the chunks in principle. A downside of this strategy is that only those algorithms (and R functions in general) can be performed that are explicitly designed to deal with hard disc specific datatypes.

IV) Integration of higher performing programming languages like C++ or Java

The integration of high performance programming languages is another alternative. Small parts of the program are moved from R to another language to avoid bottlenecks and performance expensive procedures. The aim is to balance R's more elegant way to deal with data on the one hand and the higher performance of other languages on the other hand.

The outsourcing of code chunks from R to another language can easily be hidden in functions. In this case, proficiency in other programming languages is mandatory for the developers, but not for the users of these functions.

rJava, a connection package of R and Java, is an example of this kind. Many R-packages take advantage of it, mostly invisible for the users.

V) Scope of big data analysis using R

For statistical data analysis, R is an open source software platform. Largely because of its open source nature, R is speedily adopted by statistics departments in universities around the world, attracted by its extensible nature as a platform for academic research. [1]Free in cost surely played a role as well. And it wasn't long before researchers in data science, statistics and machine learning started to publish papers in academic journals along with R code applying their new methods. R builds this process very easily and anyone can produce an R package to CRAN that stands for Comprehensive R Archive Network and make it available to everyone. An excellent open-source interactive development environment has been created by R Studio for the R language, further boosting the productivity of R users everywhere. Google, Ford, Twitter, US National Weather Service, The Rockefeller Institute of Government, The Human Rights Data Analysis Group makes use of R.

6. CONCLUSION

To create a powerful and reliable statistical model, data transformation, evaluation of multiple model options, and visualizing the results are essential. This is the reason why the R language has proven so popular: its interactive language uplifts exploration, clarification and presentation. Revolution R Enterprise gives the big-data support and speed to allow the data scientist to repeat through this process quickly.

REFERENCES

[1] http://www.r-statistics.com/tag/hadley-wickham/

[2] http://www.infoworld.com/article/2940864/application-development/r-programming-language-statistical-data-analysis.html

[3] http://spectrum.ieee.org/computing/software/the-2015-top-ten-programming-languages

[4] http://www.analytics-tools.com/2012/04/r-basics-introduction-to-r-analytics.html

[5] http://blog.revolutionanalytics.com/

[6] http://www.r-bloggers.com/handling-large-datasets-in-r/

[7] http://www.analytics-tools.com/2012/04/r-basics-introduction-to-r-analytics.html

[8] http://data.vanderbilt.edu/~hornerj/brew/useR2007.rhtml

[9] http://blog.ukdataservice.ac.uk/the-power-of-r-methods-for-processing-big-data/

[10] http://bigdatauniversity.com/moodle/course/view.php?id=522

[11] http://aisel.aisnet.org/cgi/viewcontent.cgi?article=3785&context=cais

[12] http://www.revolutionanalytics.com/what-r

[13] http://blog.revolutionanalytics.com/2013/12/tips-on-computing-with-big-data-in-r.html

[14] http://blog.ukdataservice.ac.uk/the-power-of-r-methods-for-processing-big-data/

[15] http://www.odbms.org/blog/2013/02/on-big-data-analytics-interview-with-david-smith/

[16] http://www.slideshare.net/bytemining/r-hpc

[17] https://www.pluralsight.com/blog/software-development/r-programming-language

[18] http://spectrum.ieee.org/computing/software/the-2015-top-ten-programming-languages

[19] http://www.stat.yale.edu/~mjk56/temp/bigmemory-vignette.pdf

[20] https://rpubs.com/msundar/large_data_analysis

[21] http://r.cs.purdue.edu/pub/ecoop12.pdf

[22] http://www.inside-r.org/why-use-r

[23] http://www.unt.edu/rss/R_Programming_Notes.pdf

[24] http://www.cyclismo.org/tutorial/R/input.html

[25] http://www.rosettacode.org/wiki/Category:R

[26] https://cran.r-project.org/doc/contrib/Lam-IntroductionToR_LHL.pdf