Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 46 (2015) 582 – 588

Procedia Computer Science

International Conference on Information and Communication Technologies (ICICT 2014)

Quantitative Evaluation of Big Data Categorical Variables through R

Rajiv Pandey [a]*, Manoj Dhoundiyal[b]

[a]Amity Institute of Information Technology, Amity University, Lucknow Campus, Lucknow-226028, India
[b]IT-Department, Amity University, Lucknow Campus, Lucknow-226028, India

Abstract

Big Data is the buzz word doing rounds in all areas of human existence be medical, social networks, research, it has also made inroads to education. The large size and complexity of datasets in Big Data need specialized statistical tools for analysis where R can come handy. The Categorical component of any data set can be quantified using limited representations, but evaluating it with respect to the quantitative variables return a larger set of statistical inferences. This paper explores the analysis of categorical and quantitative variables scalable to Big Data in education using a contemporary statistical tool R. R provides multiple dimensions to statistical analysis of dataset, this paper however explores the statistical inference rendered using the Box Plot feature through summary measures of the dataset. These statistical inferences can be used to train a Machine for predictions and classification under a certain category.

Peer-review under responsibility of organizing committee of the International Conference on Information and Communication Technologies (ICICT 2014)

Keywords: Categorical variables; Qantitative Variables; R; Boxplot; Big Data

1. Introduction

Big Data is a collection of both structured and unstructured data, originating from multiple sources that exhibit categorical and quantitative component, where the size of data is growing manifold with each passing second. IOT

* Corresponding author. Tel.: +919838988927
E-mail address: rpandey@lko.amity.edu

Peer-review under responsibility of organizing committee of the International Conference on Information and Communication Technologies (ICICT 2014)
doi:10.1016/j.procs.2015.02.097

(Internet of Things) is significantly contributing to the task of generating data. The devices from Mobile phone to Cyber physical Systems are all contributing to the volumes and adding to the size. The social networking application like Facebook and Twitter are a major source of data generation. The trillions of data generated from diverse nature of devices leading to diverse and complex data set is being difficult to store and process.

Processing of such large amount of data and extracting valuable information out of them in time bound fashion is one of the biggest issues while handling Big Data applications which explore a complex and colossal dataset and extract useful information or knowledge for future actions[1] .

The size of data being colossal any analysis on it is huge challenge, however the categorical and quantitative features of the dataset add to the problem. The data size is large and knowledge extraction challenges are being overcome by various tools. R is one such tool that helps in data analytics.

R is an open source programming language and software environment which supports statistical computing and visualization[5].

This paper explores the analytical inference that can be drawn out of the dataset by relating categorical variable with quantitative information using R when applied to education domain. The paper is divided into four sections 1 being the Introduction, Section 2 : "Big Data: features and dimensions" help us to list the aspects which differentiate it from any other dataset/databases, this section also list the sources of data contributing to Big Data in Education, Section 3: "Features of Categorical and Quantitative Variables", help us to understand which tool to be deployed and what summary features can be used to draw inferences on such a huge data set, Section 4 of the paper "Statistical inference of categorical variables" describes the various statistical measures and relates them to categorical separation and studies the extent to which they are influenced by the extreme values of the dataset or outliers, and the size of spread across the measures .

2. Big Data: Features and Dimensions

2.1. Big Data characteristics

Big data was first conceptualized as a data set during 2012 US presidential elections when the debate between President Barak Obama and Governor Mitt Romney generated millions of tweets every hours. The analysis or generation of this size was not thought of and the need was felt to analyze the trends at a given point of time to study the likings of the voter base. Facebook can be thought of another Big data application where millions of photographs and images are uploaded, handling this size of data needs different set of technologies w.r.t Storing, Mining and Analyzing thereby generating meaningful predictions.

Big Data technologies are being researched worldwide and most researchers have agreed upon certain features of the big data.

Big Data starts with large-volume, heterogeneous and autonomous sources that are distributed and controlled centrally, which seeks to explore complex and evolving relationships among data[2], HACE theorem.

Big Data is generally characterized by the three V's i.e. Volume, Variety, Velocity, but of late the Big Data has demonstrated certain additional dimensions.

The Features of Big Data[3] can be summarized as:

- Volume: size growing many folds with each minute passing by.
- Variety: Multi format Data.
- Velocity: The dynamics at which it is multiplying.
- Veracity: Confirming to facts and the fear in the minds of decision makers.
- Complexity: Multi device point of generation. Their integration is a challenge.

Researchers have defined, Big Data in the C3 Space[4] so that the Storage, Mining, Data analytics and Machine Learning aspects can be mathematically modeled and statistically analyzed for better utilization. The three C are

- Cardinality: Defines the number of records.
- Continuity: including the characteristics of data representation and data size.
- Complexity : includes three dimensions as
    - Large variety of data types.
    - High dimensional dataset.
    - Demand of high speed data processing.

The size and complexities involved with Big Data is a matter of concern but if effective data analytical tools deployed can be advantageous.

2.2. Big Data in Education

The education has also transformed in the current era. It has turned from offline/Indoors to online. The student data size has extended from 100 to millions. The size of data generated in the education domain has grown many folds; the education has changed from chalk and talk to mobiles and PDA's. The needs of the online students have changed w.r.t time, language and scope of the subject. It therefore is of significant concern to handle their queries and provide tailor made curriculum and suggestion to opt for suitable courses.

Big Data sources[3] in education are:

- Documents in non-electronic form,
- Data from information systems,
- Logs from university servers,
- Opinions from social networks
- Data from public education portals.

The sources being numerous the input and size of data becomes really Big, therefore all the aspects of Big Data, Machine Learning and predictions shall apply to education domain like it may apply to any other area. The size and dimension being so diverse it calls for a tool that can handle the dataset and generate valuable outcomes.

The data analytics in education is suggested using a Big Data analytical tool R[6] in the subsequent sections of the paper, primarily focusing on the quantitative and category variables.

3. Features of Categorical and Quantitative Variables

Categorical variables and quantitative variables both play a significant role in the analysis of any data set there are multiple statistical measures in respect to quantitative variables but only a few exist when it comes to categorical variables.

"*One approach is to represent the categories with numerical values (quantification) prior to visualization using methods for numerical data*"[7].

True inference is however rendered when categorical and quantitative variables are both studied in relation to each other. This section lists the features of each and performs an interrelated study in the subsequent sections

3.1. Categorical Variables

Features that are exhibited by categorical variables are:

- A count of various categories of the data set.
- A graphic visualization like Bar Chart, pie chart can be generated.
- A relative frequency measure can be obtained.

3.2. Quantitative Variables

These variables provide a larger set of statistical measures which provide a detailed inference related to the data set under considerations, like

- Minimum and Maximum value
- Quartiles
- Mean, Median and Mode
- Standard Deviation
- Box plot and scatter plot analysis.
- Histogram representation
- Extreme values
- Skew measures or Symmetric/non Symmetric behavior

The following section deliberates the relationship between both quantitative and categorical variables and draw valuable inference on a data set that may be extended to BIG Data in any area, but our context is of Education domain.

4. Statistical inference of categorical variables

The comparative analysis of categorical and quantitative variables is carried out on a dataset of approximately 300 students of two courses i.e. B-Tech-CS and B-Tech-EC. The categorical variable being Course (B-Tech-CS and B-Tech-EC courses) while quantitative variables being marks obtained by them in an Aptitude test (Max marks 70). The tool used for various analytical inferences is R, an open source data analytical tool.

The output summary measures of, minimum value, maximum value, value at 1st quartile, value at 3rd quartile, median and mean, for both the courses Fig. 1 is reproduced and further compared and analyzed along with their scatter plots and a side by side box plot. The summary measures Fig. 1 are generated after the entire dataset is read into R.

```
> summary(dataCS[,2])
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   7.00   32.00   39.00   38.04   46.00   60.00
> summary(dataEC[,2])
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  27.00   40.00   45.00   44.17   48.00   54.00
>
```

Fig. 1. Summary measure of the categorical variable

The summary measures provide certain observations i.e minimum and maximum value for B-Tech-CS is 7 and 60 while that for B-Tech-EC is 27 and 54 respectively. Thus the range of marks/statistical spread of marks obtained by students in B-Tech-CS are larger than the range for B-Tech-EC students. This shows larger difference in marks obtained (hence aptitude) between students with maximum and minimum marks for those enrolled in B-Tech-CS course. The same is shown by drawing scatter plot for students in both the courses with marks sorted in ascending order.

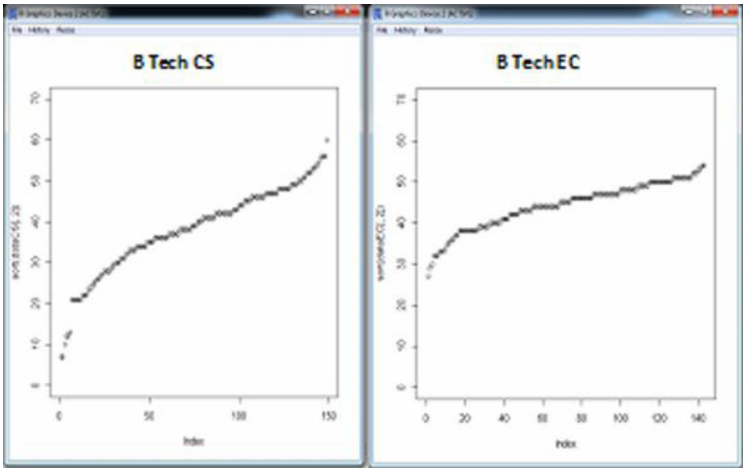

Fig. 2. Scatter plot for B Tech CS and B Tech EC with marks arranged in ascending order

The IQR or Inter Quartile Range (value at 3rd quartile – value at 1st quartile) for B-Tech-CS (46-32) is 14 and for B-Tech-EC (48-40) is 8.IQR is an important summary measure that helps to study the spread of data items and the range that they lie in. It is observed that large numbers of students of B-Tech-EC are clustered around the median of 45. This concentration may be observed in scatter plot Fig. 3.

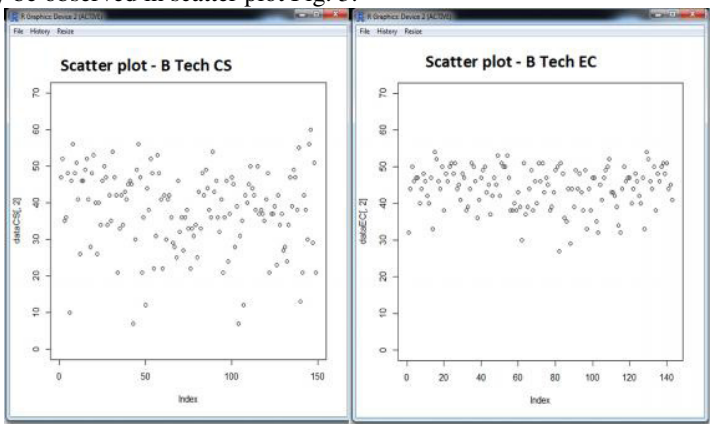

Fig. 3. Scatter plot for B Tech CS and B Tech EC

A much visible comparative study for these summarized values may be depicted with a side by side box plot for both the courses which leads to detailed observations about the categorical and quantities variables.

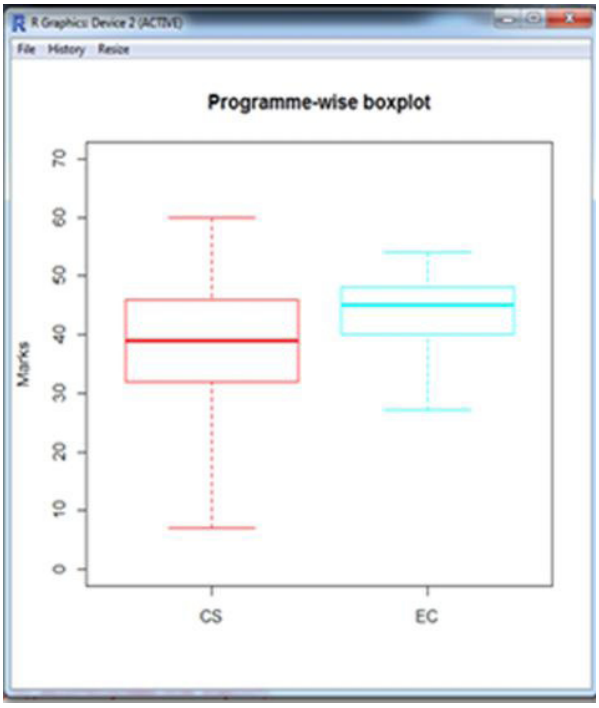

Fig. 4. Comparing Box Plot for B Tech CS and B Tech EC

Through Fig. 4 boxplot comparison we are coerced to conclude that students enrolled in B-Tech-EC are of approximately same aptitude and students enrolled in B-Tech-CS course have varying degree of aptitude or the spread of category B-Tech-CS is large thus we can predict through various machine learning algorithm that a student of certain aptitude grade will opt for B-Tech-CS or B-Tech-EC

The analytical inferences resulting from the above description can be summarized as

- Min. and Max. summary measures in Fig. 1 are resulting to the Range of spread of the data item in the two categories. Range = Max-Min i.e. 60-07= 53 for B-Tech-CS and 54-27=27 for B-Tech-EC.

- Inter Quartile Range (IQR) a measure resulting from the subtraction of 3rd quartile and the 1st quartile, 46-32= 14 for B-Tech-CS and 48-40=08 for B-Tech-EC. We can thus infer that the most data items of B-Tech-EC are grouped over a narrow margin of 08 marks, or the students are nearly of the same aptitude level.

- The B-Tech-CS scatter plot in Fig. 2 shows the outliers at the lower end across a value of 10. The outliers can sometimes make a tremendous impact on the overall analysis of the data set and need to be removed using trimmed mean feature of data analysis.

- The whiskers in the box plot of B-Tech-CS as in Fig. 4 extends over 7 to 60 thereby signifying that the spread of data is large as compared to B-Tech-EC

It is thus signified that the various statistical measures along with the graphical representations provided by R can be useful in generating data features that can serve human for better decision making and also be used as dataset for machine training and testing thereby enhancing predictability in machine learned classifications.

5. Conclusion

The statistical representation of the five summary measures of any Big Data dataset and its relationship with the categorical dimension has been explored in the paper. The mapping of the statistical measures and category has been carried out in relation to the Boxplot, Scatterplot representation. The extent of influence of outliers to the statistical measure of the median, the spread across the median and the extent of whiskers have been listed.

Even though the analysis is performed on a small data set it can be conveniently extended to a Big Data environment in Education by transforming unstructured big data into structured format through appropriate rules. The analysis resulting in the findings can be used to train, machines for classification problems in a data analytics environment. Thus an information processing activity shall scale to decision making

References

1. Rajaraman A, Ullman J. *Mining of massive Data Sets*, Cambridge Univ. Press, 2011.
2. Xindong wu, Xingquan Zhu,Gong-Qing, Wel Ding. Data Mining with Big Data, *IEEE Transactions on knowledge and data engineering*,Vol 26,No.1,January 2014.
3. Peter Michalik, Jan Stofa, Iveta Zolotova. Concept Definition for Big Data Architecture in the Education System, *IEEE 12th international Synposium on Applied Machine Intelligence and Informatics*,January 2014.
4. Shan Suthaharan. Big Data Classification:problems and challenges in Network Intrusion Prediction with Machine Learning, *Performace Evaluation Review*,Vol. 41, No 4 March 2014 .
5. Wei Fan, Albert Bifet. Mining Big Data: Current Status and Forecast to the Future, *SIGKDD Explorations* ,Volume 14 issue 2.
6. Venablles WN, Smith DM, the R Core Team. An Introduction to R, *Notes on R: A Programming Environment for Data Analysis and Graphics*,2013.
7. Sara Johansson Fernstad, Jimmy Johansson. A Task Based Performance Evaluation of Visualization Approaches for Categorical Data Analysis, IEEE DOI 10.1109/IV.2011.92, 2011