Chapter 43
On the Effectiveness of Hybrid Canopy With Hoeffding Adaptive Naive Bayes Trees: Distributed Data Mining for Big Data Analytics

Mrutyunjaya Panda
Utkal University, India

ABSTRACT

The Big Data, due to its complicated and diverse nature, poses a lot of challenges for extracting meaningful observations. This sought smart and efficient algorithms that can deal with computational complexity along with memory constraints out of their iterative behavior. This issue may be solved by using parallel computing techniques, where a single machine or a multiple machine can perform the work simultaneously, dividing the problem into sub problems and assigning some private memory to each sub problems. Clustering analysis are found to be useful in handling such a huge data in the recent past. Even though, there are many investigations in Big data analysis are on, still, to solve this issue, Canopy and K-Means++ clustering are used for processing the large-scale data in shorter amount of time with no memory constraints. In order to find the suitability of the approach, several data sets are considered ranging from small to very large ones having diverse filed of applications. The experimental results opine that the proposed approach is fast and accurate.

1. INTRODUCTION

In this era of big data, there is a tremendous change in the analysis of traditional data mining techniques. As per (Gartner, 2015; and Loney, 2001) big data is an abstract concept that can be described in terms of 3 V's as high volume, velocity and variety of information along with recent 2 V's as: variability and value to obtain meaningful insights for effective, efficient and cost effective yet innovative decision making. Even though there are different opinions what a big data is?, still, general perception about

DOI: 10.4018/978-1-5225-7501-6.ch043

big data stems from the fact which cannot be perceived, acquired, managed or processes by traditional methods within an acceptable time limit. However, this is contradicted by many stating that this notion of big data is assumed to benefit the giant software industries while totally neglecting the basic requirement of a typical user. Big data is often exaggerated and for most of the users, the physical size of the data is rarely an issue, argued by the authors (Crotty et al.2015). They also stated that big companies like: Facebook, Google, Yahoo! etc. use big data analytic job hardly above 100GB (Rowstron, Narayan, Donnelly et al., 2012) and Cloudera customers within few TB of data (Chen, Alspaugh and Katz, 2012). Even though Big Data is a hot area of research, it is not away from controversies (Boyd and Crawford, 2012). Some of them includes, but not restricted to (W. Fan and A. Bifet, 2012): (i) Big data analytics is same as traditional data analytics, as data growing is continuous over time, (ii) data recency is most important than its size during real time analysis, (iii) Big data are not always the best data for analysis. It may contain some noise as Twitter data cannot be considered as the data for counting global population and finally, most importantly, the big data management companies try to sell their products by creating a hype in the minds of the user which in due course of time may not be the best programming choice in mapreduce or Hadoop based systems. Despite so many contradiction in the definition of big data, the research in this direction must go on by dealing extraction, usefulness and transformation of a bag of data to 'Big data' (Chen, Mao & Liu, 2014).

1.1. Motivation

The main motivation begins here by considering the limits of hardware that can be provided with various machine learning algorithms; large dataset may be reduced to an extent that will be useful for the algorithm to perform efficiently, both in terms of speed and accuracy. Data reduction techniques are considered to be better in comparison to the full iteration on the big dataset, so that data is reduced with no loss of information and most importantly, after reduction, the big dataset would fit into the memory and as a result of which, the performance of the algorithm is enhanced to a greater extent. Clustering algorithm in this scenario fits well to analyse the dataset in an effective manner.

1.2. Contributions

In this research work, the contributions are as follows:

- Efficient Data subset selection using Canopy clustering for obtaining the initial seeds of canopy centers, so that the samples can fit to the CPU cache size, for faster learning. This is considered to be an important step to divide the whole data into small cluster sizes and to make the Big Data processing more efficient and scalable.
- K-means++ clustering and Hoeffding trees are used as a classification technique to the data, after parameter setting by initialization of Canopy cluster centers is done apriori in the 1st step. This assures of boosting the performance of the proposed Big data analysis, in terms of accuracy and speed.
- In this paper, we evaluate the performance on different real world dataset such as: Agricultural dataset (Mushroom, Soybean, Eucalyptus, grub-damage, Pasture Production, Squash stored, white clover), Airlines dataset and Google Cluster dataset that differ in terms of volume, variety, velocity etc., for implementation.