Chapter 41
Essentiality of Machine Learning Algorithms for Big Data Computation

Manjunath Thimmasandra Narayanapppa
BMS Institute of Technology, India

T. P. Puneeth Kumar
Acharya Institute of Technology, India

Ravindra S. Hegadi
Solapur University, India

ABSTRACT

Recent technological advancements have led to generation of huge volume of data from distinctive domains (scientific sensors, health care, user-generated data, finical companies and internet and supply chain systems) over the past decade. To capture the meaning of this emerging trend the term big data was coined. In addition to its huge volume, big data also exhibits several unique characteristics as compared with traditional data. For instance, big data is generally unstructured and require more real-time analysis. This development calls for new system platforms for data acquisition, storage, transmission and large-scale data processing mechanisms. In recent years analytics industries interest expanding towards the big data analytics to uncover potentials concealed in big data, such as hidden patterns or unknown correlations. The main goal of this chapter is to explore the importance of machine learning algorithms and computational environment including hardware and software that is required to perform analytics on big data.

INTRODUCTION

Every day, 2.5 quintillion bytes of data are created and 90 percent of the data in the world today were produced within the past two years (IBM, 2012). Our capability for data generation has never been so powerful and enormous ever since the invention of the information technology. As another example, in 2012, the first presidential debate between President Barack Obama and Governor Mitt Romney generated

DOI: 10.4018/978-1-4666-9840-6.ch041

more than 10 million tweets in 2 hours (Twitter, 2012). Among all these tweets, the specific moments that generated the most discussions revealed the public interests, such as the discussions about vouchers and Medicare. Such online discussions provide a new means to sense the public interests and generate feedback in real- time, and are mostly appealing compared to standard media, such as TV broadcasting, newspapers or radio. Another example is Flickr, a picture sharing site, which receives on an average 1.83 million photos (Michel, 2015). Assuming the size of each photo is 2 megabytes (MB), this requires 3.6 terabytes (TB) of storage disk every single day. In fact, as an old saying states: "a picture is speaks a thousand words," the billions of pictures collected by Flicker are a treasure tank for us to explore the human society, public affairs, social events, disasters, and so on, only if we have the powerful technology to harness the enormous amount of data. The above examples show the rise of Big Data applications where data collection has grown tremendously and is beyond the ability of commonly used software tools to acquire, manage, and process within an "acceptable elapsed time." An essential challenge facing by applications of Big Data is to explore the large volumes of data and extract useful information or knowledge for future actions (Rajaraman & Ullman, 2011).

Machine learning is a branch of artificial intelligence that allows us to make our application intelligent without being explicitly programmed. Machine learning concepts are used to enable applications to take a decision from the available datasets. A combination of machine learning and data mining can be used to develop various applications such as spam mail detectors, self-driven cars, face recognition, speech recognition, and online transactional fraud-activity detection. There are many popular organizations that are using machine-learning algorithms to make their service or product understand the need of their users and provide services as per their behavior. Google has its intelligent web search engine, which provides a number one search, spam classification in Google Mail, news labeling in Google News, and Amazon for recommender systems. There are many open source frameworks available for developing these types of applications/frameworks, such as R, Python, Apache Mahout, and Weka (Han Hu, Wen, Chua, Xuelong Li, n.d).

In the basic computational model of CPU and memory, the algorithms runs on the CPU and access the data that is in the memory, its need to bring in the data from disk into memory, but once the data is in memory, and the algorithm runs in the data that is on memory. This is the familiar model considered to implement all kinds of algorithms such as machine learning, statistics etc. wherever the data is so big, that it cannot fit into memory at the same time. That's where data mining comes in. In traditional data mining algorithms, since the data is a big, only portion of the data bring into memory at a time and process the data in batches, finally writes the results back to disk. But sometimes even this is not sufficient. Now if you take ten billion webpages, each of 20 KB, you have, total dataset size of 200 TB. Now, when you have 200 TB, let us assume that by using the traditional computational model, traditional data mining model. And all this data is stored on a single disk, and we have read tend to be processed inside a CPU. Now the fundamental limitation here is the data bandwidth between the disk and the CPU. The data has to be read from the disk into the CPU, and read bandwidth for most modern SATA disk is around 50MB a second, so we can read data at 50MB a second, it means its takes 4 million seconds that is 46 days. To do something useful with the data, it's going to take even longer time. Such a long time is unacceptable, we need a better solution to read and process the Big Data. One of the solutions for this kind of problem is to split the data into chunks on multiple disks and CPUs. Now read the chunks of the data from the multiple disks and process it in parallel in multiple CPUs. That will cut down read and process time by a lot. For example, if you had a 1,000 disks and CPUs, 4 million seconds come down to 4,000 seconds. This is the fundamental idea behind the idea of cluster computing.