Chapter 24
Big Data Mining Based on Computational Intelligence and Fuzzy Clustering

Usman Akhtar
Air University – Multan, Pakistan

Mehdi Hassan
Air University – Multan, Pakistan

ABSTRACT

The availability of a huge amount of heterogeneous data from different sources to the Internet has been termed as the problem of Big Data. Clustering is widely used as a knowledge discovery tool that separate the data into manageable parts. There is a need of clustering algorithms that scale on big databases. In this chapter we have explored various schemes that have been used to tackle the big databases. Statistical features have been extracted and most important and relevant features have been extracted from the given dataset. Reduce and irrelevant features have been eliminated and most important features have been selected by genetic algorithms (GA). Clustering with reduced feature sets requires lower computational time and resources. Experiments have been performed at standard datasets and results indicate that the proposed scheme based clustering offers high clustering accuracy. To check the clustering quality various quality measures have been computed and it has been observed that the proposed methodology results improved significantly. It has been observed that the proposed technique offers high quality clustering.

1. INTRODUCTION

The era of petabyte (10^{15}) has almost gone, leaving us to confront of zettabytes (10^{21}) era. Technology uprising has been facilitating millions of people all around the globe to generate tremendous amount of data via ever-increased used of variety of digital technology that generate continuous streams of data. A recent survey by New Vantage Partner (Davenport, 2013) reports that "It is about variety, not volume", but many people still believe that the issue with big data is either scale or volume. Big data involves variety of data forms like text, images, videos, and sounds etc. It is estimated that nearly twenty percent of

DOI: 10.4018/978-1-5225-7501-6.ch024

data available to enterprises is structured in nature and other eighty percent is unstructured data (Judith Hurwitz, April 2013).

From the data mining perspective, big data has opened many new challenges and opportunities. Even big data involves hidden knowledge and insights; it brings many new issues to extract these hidden insight patterns. The traditional processes of knowledge discovery processes are not well suited for big data. In general, the existing data mining techniques encounter difficulties when they are require handling heterogeneity, volume, privacy, and accuracy. KDD is a knowledge discovery data process of unveiling hidden knowledge and insights from a large volume of data (Fayyad, Piatetsky-Shapiro, & Smyth, 1996), when it comes to extract useful information from big data then it became a most interesting and challenging step. Another problem that arises in the data mining algorithms is to inadequate scalability of the algorithms that do not matches the 3-Vs (variety, velocity, and volume) of the emerging big data. Big data not only brings new challenges, but it also brings new opportunities. Big data would become a useless monster if we don't have the right tools to harness its "wildness"(Che, Safran, & Peng, 2013). Current data mining techniques are not ready to meet new challenges of big data.

In data mining, the conventional data mining algorithms have difficulties in handling the challenges drawn by the unstructured data which is often vague and uncertain. For the pattern recognition and data mining, clustering is mostly used to search in very large databases. So, there is a need of clustering algorithms that used to search large datasets (Havens, Bezdek, Leckie, Hall, & Palaniswami, 2012).In clustering each group share some similarity and clustering is a form of data analysis.

2. BIG DATA CHARACTERISTICS

Big Data involves large datasets which are complex and cannot easily analyzed, interprets, and processed further (Sagiroglu & Sinanc, 2013). Big Data concern large-volume, complex growing datasets with multiple, autonomous sources of data generation (Xindong, Xingquan, Gong-Qing, & Wei, 2014). Big Data is usually in large volume and heterogeneous characteristics of Big Data that is extremely challenging for discovering useful information. The main attributes of Big Data are volume, velocity, and variety. Volume is one of the main features of Big Data that are massive in scale and usually generated in every second. It is a scale characteristic. Another attributes are velocity or speed such as social media data. Variety is one of the main attribute of big data that is heterogeneous in nature. It includes different types such as text data, images and multi-dimensional arrays data.

3. DATA MINING CHALLENGES WITH BIG DATA

The goals of the data mining techniques go beyond extracting requested information or even hidden patterns and it must deal with heterogeneity, scalability, and accuracy. There is a need for designing and implementing large scale machine learning and data mining algorithms which accomplish the processing of very large scale data. There are two main challenging areas for big data mining. These areas are computing platform problem, and big data mining algorithms problem.

Big Data mining platform is a major challenge on data accessing and processing. Big Data often stored at different locations and volume of data is continuously growing. An effective computing platform is highly desirable. Typical data mining algorithms load all the data into the main memory, this however a