Chapter 108
Advanced Dimensionality Reduction Method for Big Data

Sufal Das
North-Eastern Hill University, India

Hemanta Kumar Kalita
North-Eastern Hill University, India

ABSTRACT

The growing glut of data in the worlds of science, business and government create an urgent need for consideration of big data. Big data is a term that describes large volumes of high velocity, complex and variable data that require advanced techniques and technologies to enable the capture, storage, distribution, management, and analysis of the information. Big data challenge is becoming one of the most exciting opportunities for the next years. Data mining algorithms like association rule mining perform an exhaustive search to find all rules satisfying some constraints. it is clear that it is difficult to identify the most effective rule from big data. A novel method for feature selection and extraction has been introduced for big data using genetic algorithm. Dimensionality reduction can be considered a problem of global combinatorial optimization in machine learning, which reduces the number of features, removes irrelevant, noisy and redundant data, to obtain the accuracy and saves the computation time and simplifies the result. A genetic algorithm was developed based approach utilizing a feedback linkage between feature selection and association rule using MapReduce for big data.

1. INTRODUCTION

Information is gathered almost everywhere in our everyday lives. Industries are generating huge amount of digital data as they go about their business and interactions with individuals. Big data is being presented through social media sites, smart phones, and other consumer devices including PCs and laptops which are being used by billons of individuals around the world.

Big data refers to very large datasets whose size is beyond the ability of typical software tools to gather, store, process, manage, and analyze (Gopalkrishnan, V., Steier, D., Lewis, H., & Guszcza, J. 2012). Big dataset needs to be in order to be considered big data i.e., we don't define big data in terms

DOI: 10.4018/978-1-4666-9840-6.ch108

of being larger than a certain number of terabytes (thousands of gigabytes) only. It is assumed that, as technology advances over time, the size of datasets that qualify as big data will also increase. Also the definition can vary by different fields, depending on what kinds of software tools are commonly available and what sizes of datasets are common in a particular company. Big data in many sectors today will range from a few dozen terabytes to multiple petabytes (thousands of terabytes).

For increasing of the amount of data in the field of medicals, management, genomics, communication, biology, environmental research and many others, it has become difficult to built, study patterns, relations within such large data.

Big data can be deliberated using the 4 V's: Volume, Velocity, Variety, and Veracity (Boyd, D., & Crawford, K. 2012).

- **Volume:** Everyday large volume of data is being collected from social media sites, smart phones, and other consumer devices. Too much volume is a storage issue, as well as too much data is also a massive analysis issue as traditional database system fails.
- **Velocity:** Velocity means both how fast data is being produced and how fast the data must be processed to meet demand. as large volume of data is being generated, it is very important to process with synchronization for a system.
- **Veracity:** Industry leaders don't want to share the information which they use to make decisions. Establishing trust in big data presents a huge challenge as the variety and number of sources grows.
- **Variety:** Translating large volumes of transactional information into decisions is a major concern while big data is considered. Now there are many types of information to analyze, mainly coming from social media and communication devices. Variety includes structured data like tabular data (databases), transactions etc. and unstructured and semi-structured data like hierarchical data, documents, e-mail, video, images, audio etc.

It would be difficult and time consuming for handling big data and we have to follow certain algorithm and method to analyze the data, find an appropriate classification among them. The standard data analysis method such as probing, clustering, factorial, analysis needs to be extended to get the information and extract new knowledge.

Feature selection is broad and spread across many fields, including document classification, data mining, object recognition, biometrics, remote sensing and computer vision. It is relevant to any job where the number of features or attributes is bigger than the number of training examples, or excessively huge to be computationally attainable. Feature selection is likewise identified with four different areas of exploration: dimensionality reduction, space partitioning, and feature extraction and decision tree. Most of the data includes irrelevant, redundant, or noisy features. Feature selection reduces the number of features, removes irrelevant, redundant, or noisy features, and brings about palpable effects on applications by speeding up a data mining algorithm, improving learning accuracy, and leading to better model comprehensibility.

Data mining algorithms like association rule mining (ARM) (Agrawal, R., Imieliński, T., & Swami, A. 1993) perform a comprehensive pursuit to discover rules satisfying some constraints. Hence, the number of discovered rules from database can be very large. Taking into account the prior works, it is clear that to identify the most effective rule is difficult. Therefore, in many applications, learning may not work well before removing the unwanted features as the size of the dataset is so large.