Big Data Research 2 (2015) 87–93

Contents lists available at ScienceDirect

Big Data Research

www.elsevier.com/locate/bdr

Efficient Machine Learning for Big Data: A Review ☆

Omar Y. Al-Jarrah [a], Paul D. Yoo [b,*], Sami Muhaidat [c], George K. Karagiannidis [a,d], Kamal Taha [a]

[a] *Khalifa University, Abu-Dhabi, United Arab Emirates*
[b] *Data Science Institute, Bournemouth University, UK*
[c] *University of Surrey, Guildford, UK*
[d] *Aristotle University of Thessaloniki, Thessaloniki, Greece*

ARTICLE INFO

Article history:
Received 21 December 2014
Received in revised form 23 March 2015
Accepted 7 April 2015
Available online 20 April 2015

Keywords:
Big data
Green computing
Efficient machine learning
Computational modeling

ABSTRACT

With the emerging technologies and all associated devices, it is predicted that massive amount of data will be created in the next few years – in fact, as much as 90% of current data were created in the last couple of years – a trend that will continue for the foreseeable future. Sustainable computing studies the process by which computer engineer/scientist designs computers and associated subsystems efficiently and effectively with minimal impact on the environment. However, current intelligent machine-learning systems are performance driven – the focus is on the predictive/classification accuracy, based on known properties learned from the training samples. For instance, most machine-learning-based nonparametric models are known to require high computational cost in order to find the global optima. With the learning task in a large dataset, the number of hidden nodes within the network will therefore increase significantly, which eventually leads to an exponential rise in computational complexity. This paper thus reviews the theoretical and experimental data-modeling literature, in large-scale data-intensive fields, relating to: (1) model efficiency, including computational requirements in learning, and data-intensive areas' structure and design, and introduces (2) new algorithmic approaches with the least memory requirements and processing to minimize computational cost, while maintaining/improving its predictive/classification accuracy and stability.

1. Introduction

Today, it's no surprise that reducing energy costs is one of the top priorities for many energy-related businesses. The global information and communications technology (ICT) industry that pumps out around 830 Mt carbon dioxide (CO_2) emission accounts for approximately 2 percent of the global CO_2 emissions [1]. ICT giants are constantly installing more servers so as to expand their capacity. The number of server computers in data centers has increased sixfold to 30 million in the last decade, and each server draws far more electricity than its earlier models [2]. The aggregate electricity use for servers had doubled between the years 2000 and 2005 period, most of which came from businesses installing large numbers of new servers [3]. This increase in energy consumption consequently results in higher carbon dioxide emissions, and hence causing an impact on the environment. Furthermore, most of these businesses, especially in an uncertain economic climate are placed under the pressure to reduce their energy expenditure in order to remain competitive in the market [4].

With the emerging of new technologies and all associated devices, it is predicted that there will be as much data created as was created in the entire history of planet Earth [5]. Given the unprecedented amount of data that will be produced, collected and stored in the coming years, one of the technology industry's great challenges is how to benefit from it. During the past decade, mathematical intelligent machine-learning systems have been widely adopted in a number of massive and complex data-intensive fields such as astronomy, biology, climatology, medicine, finance and economy. However, current intelligent machine-learning-based systems are not inherently efficient or scalable enough to deal with large volume of data. For example, for many years, it is known that most non-parametric and model-free approaches require high computational cost to find the global optima. With high-dimensional data, their good data fitting capacity not only makes them more susceptible to the generalization problem but leads to an exponential rise in computational complexity.

☆ This article belongs to BDA-HPC.
* Corresponding author.
E-mail address: omar.aljarrah@kustar.ac.ae (O.Y. Al-Jarrah).

http://dx.doi.org/10.1016/j.bdr.2015.04.001
2214-5796/© 2015 Elsevier Inc. All rights reserved.

Designing more accurate machine-learning systems so as to satisfy the market needs will hence lead to a higher likelihood of energy waste due to the increased computational cost.

Nowadays, there is a greater need to develop efficient intelligent models to cope with future demands that are in line with similar energy-related initiatives. Such energy-efficient-oriented data modeling is important for a number of data-intensive areas, as they affect many related industries. Designers should focus on maximum performance and minimum energy use so as to break away from the traditional' performance vs. energy-use' tradeoff, and increase the number and diversity of options available for energy-efficient modeling. However, despite the fact that there is a demand for such efficient and sustainable data modeling methods for large and complex data-intensive fields, to our best knowledge, only a few of these literatures have been proposed in the field [6,7].

This paper provides a comprehensive review of state-of-the-art sustainable/energy-efficient machine-learning literatures, including theoretical, empirical and experimental studies pertaining to the various needs and recommendations. Our objective is to introduce a new perspective for engineers, scientists, and researchers in the computer science, and green ICT domain, as well as to provide its roadmap for future research endeavors.

This paper is organized as follows. Section 2 introduces the different large-scale data-intensive areas and discusses their structure and nature, including the relation between data models and their characteristics. Section 3 discusses the issues in current intelligent data modeling for sustainability and gives recommendations. Section 4 concludes the paper.

2. Big data challenge

e-Science areas are typically data-intensive in that the quality of their results improves with both quantity and quality of data available. However, current intelligent machine-learning systems are not inherently efficient enough which ends up, in many cases, a growing fraction of this quantity data unexplored and underexploited. It is no small problem when existing methods fail to capture such data immensity. When old concepts fail to keep up with change, traditions and past experience become inadequate guide for what to do next. Effective understanding and the use of this new wealth of raw information pose a great challenge to today's green engineers/researchers. It should be noted that the scope of the review is limited to the analytical aspects of science areas using immense datasets, and the methods for reducing computational complexity in distributed or grid-computing environment are excluded.

2.1. Geo, climate and environment

There are many recent examples that can illustrate the tremendous growth in scientific data generation in the literature. It is estimated that there are thousands of wireless sensors currently in place, which generates about a gigabyte of data per sensor per day [8]. Such sensors measure and record sensory information about the natural environment at a joint spatial and temporal dimensions that has never previously been possible. This environmental information is gathered by sensors via its sensing devices that are attached to small, low-power computer systems with digital radio communications. The sensor nodes self-organize itself into a network to deliver, and perhaps process the collected data to a base station, where it can be made available to the users through the Internet. These sensors generate several petabytes of data per year and decisions need to be taken in real time as to how much data to analyze, how much to transmit for further analysis.

Besides the environmentalists, a similar challenge facing the climatologists, meteorologists, and geologists today is also making sense of the vast and continually increasing amount of data generated by the earth observation satellites, radars, and high-throughput sensor networks. The World Data Centre for Climate (WDCC) is the world-largest climate data repository, and is also known to have the largest database in the world [9]. The WDCC archives 340 terabytes of earth system model data and related observations, and 220 terabytes of data readily accessible on the web including information on climate research and anticipated climatic trends, as well as 110 terabytes (or 24,500 DVD's) worth of climate simulation data. The WDCC data is accessible by a standard web-interface (http://cera.wdc-climate.de). These data are increasingly available in many different formats and have to be incorporated correctly into the various climate change models. Timely and accurate interpretation of these data can provide advance warnings in times of severe weather changes, hence enabling corresponding action to be taken promptly so as to minimize its resulting catastrophic damage.

2.2. Bio, medicine, and health

Biological data has been produced at a phenomenal rate due to the international research effort called the Human Genome Project. It is estimated that the human genome DNA contains around 3.2 billion base (3.2 gigabase) pairs distributed among twenty-three chromosomes, which is translated to about a gigabyte of information [10]. However, when we add the gene sequence data (data on the 100,000 or so translated proteins and the 32,000,000 amino acids), the relevant data volume can easily expand to an order of about 200 gigabyte [11]. Now, by including also the X-ray/NMR spectroscopy structure determination of these proteins, the data volume will increase dramatically to several petabytes, and that is assuming only one structure per protein.

As of December 2014, the GenBank repository of nucleic acid sequences contained above 178 million entries [12] and the SWISS-PROT database (inc. both UniProtKB/Swiss-Prot, UniProtKB/TrEMBL) of protein sequences contained about 18 million entries [13,14]. On average, these databases are doubling in size in every 15 months. This is further compounded by data generated from the myriad of related projects that study gene expression, that determines the protein structures encoded by the genes, and that details how these proteins interact with one another. From that, we can begin to imagine the enormous amount and variety of information that is being produced every month.

Over the past decade, the health sector has also evolved significantly, from paper-based systems to largely paperless electronic systems. Many countries' public health systems are now providing electronic patient records with advanced medical imaging media. In fact, this has already been implemented by more than 200 American hospitals, and the days of squinting to decipher a doctor's untidy scrawl on a handwritten prescription will soon be a thing of the past in Canada and many other countries too [15].

InSiteOne is one of the leading service providers in offering data archiving, storage, and disaster-recovery solutions to the healthcare industry. Its U.S. InSiteOne's archives include almost 4 billion medical images and 60 million clinical studies, in a coverage area of about 800 clinical sites [16]. The combined annual total of its radiological images exceeds 420 million and this number is still increasing at an approximate rate of about 12% per year. There are about 35,500 radiologists currently practicing in the U.S. [17]. Each image will typically constitute several megabytes of digital data and is required to be archived for a minimum of five years. ESG (Enterprise Storage Group) forecasts medical image data in North America will grow to more than 35 percent per year and will reach nearly 2.6 million terabytes by 2014 [18]. It is also worthwhile to