International Journal of Cyber-Physical Systems
Volume 2 • Issue 2 • July-December 2020

Modern Subsampling Methods for Large-Scale Least Squares Regression

Tao Li, Institute of Statistics and Big Data, Renmin University of China, China
Cheng Meng, Institute of Statistics and Big Data, Renmin University of China, China

ABSTRACT

Subsampling methods aim to select a subsample as a surrogate for the observed sample. As a powerful technique for large-scale data analysis, various subsampling methods are developed for more effective coefficient estimation and model prediction. This review presents some cutting-edge subsampling methods based on the large-scale least squares estimation. Two major families of subsampling methods are introduced: the randomized subsampling approach and the optimal subsampling approach. The former aims to develop a more effective data-dependent sampling probability while the latter aims to select a deterministic subsample in accordance with certain optimality criteria. Real data examples are provided to compare these methods empirically, respecting both the estimation accuracy and the computing time.

KEYWORDS

Big Data, Data Reduction, Leverage Scores, Linear Model, Optimal Design, Randomization Algorithm, Statistics, Subsample

DOI: 10.4018/IJCPS.2020070101

1. INTRODUCTION

During recent decades, the rapid development of science and technologies enables researchers to collect data with unprecedented sizes and complexities. In the meanwhile, large-scale datasets are emerging in all fields of science and engineering, from academia to industry. For example, Facebook has over 1.75 billion active users who contribute to nearly 350 million photos which are uploaded to Facebook daily (Omnicoreagency.com, 2020). Consider Twitter, around 6,000 tweets are tweeted on Twitter in a second (David Sayce, 2020). In addition, viewers spent around 15 billion hours (1,712,000 years, which is still rising) on YouTube in a month, and the videos being uploaded to YouTube are at the rate of 72 hours per minute (Omni Media, 2018). These social media platforms are collecting and generating massive datasets with various types, such as text data, image data, and video data. For another example, the European Bioinformatics Institute, one of the world's largest biology-data repositories, stores nearly 160 petabytes of data and back-ups about genes, proteins, and small molecules. Moreover, such huge amount of genomics data almost doubles annually (Cook et al., 2019).

The large-scale datasets emerging from all fields provide researchers with unprecedented opportunities for data-driven decision-making and knowledge discoveries. Nevertheless, traditional statistical and machine learning algorithms may fail to analyze these data due to considerable computational burden in terms of both time and memory. The task of analyzing large-scale datasets calls for innovative, effective, and efficient methods or algorithms for addressing the new challenges due to the explosion of data.

According to Laney (2001), the big data challenges can be evaluated in three main aspects, including volume, velocity, and variety. Specifically, the volume is the size related to both the dimension and the number of observations, velocity is the interaction speed with the data, and the variety indicates various data structures. In this article, the authors mainly discuss the first scenario with a focus on the case that the number of observations n far exceeds the data dimension p.

To alleviate the computational burden caused by large n, there has been a large number of studies dedicated to developing engineering solutions. These solutions include cloud computing, designing more powerful supercomputers and parallel computing among others. More details of these methods are provided in Section 2.

Despite the effectiveness of engineering solutions, efficient statistical solutions are still in high demand, making big data analysis manageable on general-purpose personal computers. The subsampling method is a powerful technique that can be used to achieve this goal. A subsampling problem can be described as follows: given a p-dimensional sample $\left\{x_i\right\}_{i=1}^{n}$ generated from an unknown probability distribution, the goal is to take a subsample $\left\{x_i^*\right\}_{i=1}^{r}$, $r \ll n$, as a surrogate for the original sample. The selected subsample is then processed by down-streaming analysis for coefficient estimation, model prediction and statistical inference.