Algorithms for Online Subset Selection and Data Valuation in Data-Centric AI

Abstract of the thesis to be submitted in partial fulfillment of the requirements for the award of the degree of

Doctor of Philosophy

by

Soumi Das
(17CS91P08)

Under the supervision of

Dr. Sourangshu Bhattacharya

Department of Computer Science and Engineering
Indian Institute of Technology Kharagpur

July 2023

Abstract

The advent of high connectivity across the globe has led to an incoming stream of dynamic data at a high velocity. It leads to several bottlenecks related to storage, training time, redundancies, unfairness, biases, or any form of noise in data. The inconsistencies in the data result in unstable models. Therefore, it is essential to select high-quality data that can form appropriate representatives, thus helping in mitigating the limiting factors. There has been a substantial amount of work in the field of coreset selection which assumes the data to originate from some mixture of Gaussian or Poisson distributions. However, the effectiveness of the classical coreset techniques is doubtful in the deep learning setting due to the high dimensional complex data and their representations. The recent set of works in the area of subset selection focuses on only the pairwise relationships between the datapoints to identify the representatives and that of data valuation focuses on only the information obtained from the model, to compute the scores, leading to redundant selections. Besides, they are also not flexible enough to incorporate different metrics or value functions that can help in providing subsets aimed toward specific goals like generalization or robustness. In this thesis, we try to address the limitations of the existing works and develop algorithms to obtain high-quality datasets.

Considering the dynamic growth in data in the streaming setting, the primary goal of this thesis is to develop online subset selection and data valuation algorithms as a part of data-centric AI development. Earlier works on online subset selection have focused on using mainly pairwise criteria like dissimilarity as an external criterion to find out representatives. In the first work of this thesis, we develop a multi-criteria (pairwise and pointwise) based framework for filtering out representatives from the incoming streaming data. The pointwise criterion is based on the model feedback, for example, the loss of an instance, that helps to obtain subsets aimed at an end-task objective. We design a convex-based formulation for the framework and use it for selecting representatives in the task of autonomous driving applications like episode completion and semantic segmentation. We show that the proposed method can complete 100% of the episodes even after dropping 80% of the data while the baselines can complete only 30% to 70% of the episodes.

The task of quantifying the value of individual data has become a significant research domain in recent times. In the second work of the thesis, we address the problem of finding high-value subsets from the perspective of using value functions that are targeted toward a particular goal. The existing data valuation methods only consider the model information to compute the scores. They ignore the fact that the value of a training datapoint depends on the other datapoints. In contrast to these bodies of work, we design a framework that considers the interdependencies between instances (to ensure the diversity of selected instances) along with the model information encoded in the value function. We design an online alternating minimization-based algorithm for jointly learning the parameters of the selection network and machine learning model and show its efficacy on several domains of data like image, protein, text as well as synthetically generated data while outperforming the baselines by a maximum of about 20%.

While all these data valuation and subset selection (DVSS) methods have been computationally heavy, and also not flexible enough to be adaptable to other value functions, in the final work of this thesis, we propose that in order to be more effective, DVSS techniques should follow the aspects of being flexible, accurate, robust, and efficient (FARE). We develop a two-stage online sparse approximation framework that consists of two stages: (a) checkpoint selection, where an optimal set of checkpoints are selected that best approximate a given value function over the training runs, and, (b) data valuation, where the selected set of checkpoints are used for

valuating the individual datapoints. We show the aspects of accuracy and efficiency on different domain datasets like image, text, and tabular data where CheckSelect surpasses the baselines by a maximum of about 30%. For the aspects of flexibility and robustness, we show that the selected set of checkpoints obtained from the source domain help in valuating datapoints from a target domain without any need for re-training and performs 7-8% higher than the closest performing baselines.

Keywords: online subset selection, data valuation, high-quality data, multi-criteria convex framework, learnable subset selection, online sparse approximation.

Publications from the thesis

1. **Soumi Das**, Harikrishna Patibandla, Suparna Bhattacharya, Kshounis Bera, Niloy Ganguly, and Sourangshu Bhattacharya. "TMCOSS: Thresholded Multi-Criteria Online Subset Selection for Data-Efficient Autonomous Driving." In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021.

2. **Soumi Das**, Arshdeep Singh, Saptarshi Chatterjee, Suparna Bhattacharya, and Sourangshu Bhattacharya. "Finding high-value training data subset through differentiable convex programming." In Machine Learning and Knowledge Discovery in Databases. Research Track: European Conference, ECML PKDD 2021.

3. **Soumi Das**, Sayan Mandal, Ashwin Bhoyar, Madhumita Bharde, Niloy Ganguly, Suparna Bhattacharya, and Sourangshu Bhattacharya. "Multi-criteria online frame-subset selection for autonomous vehicle videos." Pattern Recognition Letters 133 (PRL) 2020.

Under review

1. **Soumi Das**, Manasvi Sagarkar, Suparna Bhattacharya, Sourangshu Bhattacharya. "Check-Select: Online Sparse Approximation based Checkpoint Selection for "FARE" Data Valuation and Subset Selection".

Publications during PhD not included in the thesis

1. Kiran Purohit, Anurag Parvathgari, **Soumi Das**, and Sourangshu Bhattacharya. "Accurate and Efficient Channel pruning via Orthogonal Matching Pursuit." In Proceedings of the Second International Conference on AI-ML Systems 2022.