Title: LEARNING ON LARGE DATASETS USING BIT-STRING TREES

Abstract:
Similarity preserving hashing finds widespread application in nearest-neighbor search. The widely used form of similarity preserving hashing is space-partitioning-based hashing. Many space partitioning-based hashing techniques generate bit codes as hash codes. Although Binary Search Trees (BSTs) can be used for storing bit codes, their size grows exponentially with code length. In practice, such a tree turns out to be highly sparse, increasing the miss-rate of nearest neighbor searches. To tackle sparsity and memory issues of BST, we first developed Compressed BST of Inverted hash tables (ComBI), a geometrically motivated compression technique for BSTs. ComBI enables fast and approximate nearest neighbor searches without a significant memory footprint over BSTs. We show, that approximate search in ComBI is competitive with an exact search algorithm in retrieving the nearest neighbors in a hamming space. On a database containing ∼80 million samples, ComBI yields an average precision of 0.90, at ∼4X - ∼296X improvements in run-time across different code lengths when compared to Multi-Index Hashing (MIH), a widely used exact search method. On a database consisting of 1 billion samples, this value of precision (0.90) is reached at ∼4X - ∼19X improvements in run-time. Next, the ComBI has been shown as a search engine for single-cell RNA sequencing (scRNA-seq) data, and its performance is compared with the state-of-the-art scRNA-seq search engine method, Cellfishing.jl, which is based on the MIH. The ComBI outperforms Cellfishing.jl in multiple accounts. The achieved speed-up in the search is around ∼2 - ∼13.

We next shift our attention to using similarity preserving hashing to build a classifier. The learned structure of hashing algorithms is suitable to be combined with a Bayes' classifier. We explored the construction of three basic space-partitioning-based hashing algorithms and identified their pros and cons. This motivated us to build a tree-based hashing classifier. We present Guided Random Forest (GRAF), a tree-based ensemble hashing classifier that realizes global partitioning by extending the idea of building oblique decision trees with localized partitioning. We show that GRAF bridges the gap between decision trees and boosting algorithms. Experiments indicate that it reduces the generalization error bound. Results on 115 benchmark datasets show that GRAF yields comparable or better results on a majority of datasets. We also build an unsupervised version of GRAF, Unsupervised GRAF (uGRAF), to perform guided hashing. The GRAF fundamentally works by generating more hyperplanes in the region of high data complexity and this phenomenon is represented by the number of planes required to classify a sample correctly. This measure can be used for importance sampling. In the next part of the thesis, this direction is explored to build a data approximator using GRAF. An extensive empirical evaluation with simulated and UCI datasets was performed to establish the theory. The proposed methodology is compared with the two state-of-the-art importance sampling algorithms. An analogy between Support Vector Machine (SVM) and the samples marked by GRAF as of high importance is also developed.

We then show that the learned neighborhood of a sample can be used to estimate the confusion around the sample in a scalable manner. We utilized uGRAF and ComBI to estimate the per-sample classifiability. An empirical evaluation of estimated values is presented. We show how per-sample classifiability can be used to estimate cancer patient survivability.

Cancer is a disease of the genome. Genomic changes resulting in cancer can be inherited, brought on by environmental carcinogens, or may result from random replication errors. Mutations continue to spread after the induction of carcinogenicity and significantly change cancer genomes. Most cancer-related somatic mutations are indistinguishable from germline variants or other non-cancerous somatic mutations, even though only a small subset of driver mutations have been identified and

characterized thus far. Thus, such overlap makes it difficult to understand many harmful but unstudied somatic mutations. The main bottleneck results from patient-to-patient variation in mutational profiles, which makes it challenging to link particular mutations with a particular disease outcome. This thesis introduces a newly developed method called Continuous Representation of Codon Switches (CRCS). This deep learning-based approach enables us to produce numerical vector representations of genetic changes, enabling a variety of machine learning-based tasks. We show how CRCS can be used in three different ways. First, we show how it can be used to find cancer-related somatic mutations without matched normal samples. Second, the suggested method makes it possible to find and study driver genes. Finally, we created a numerical representation of mutations by combining a sequence classifier with CRCS. These representations are used to score individual mutations in a tumor sample using per-sample classifiability, which was found to be predictive of patient survival in Bladder Urothelial Carcinoma (BLCA), Hepatocellular Carcinoma (HCC), and Glioblastoma Multiforme (GBM). Taken together, we propose CRCS as a valuable computational tool for analysis of the functional significance of individual cancer mutations.

Publications:

Publications related to thesis chapters:

Gupta P, Jindal A, Ahuja G, Sengupta D. A new deep learning technique reveals the exclusive functional contributions of individual cancer mutations. Journal of Biological Chemistry. 2022 Jun 24:102177.

Gupta P, Jindal A, Jayadeva, Sengupta D. ComBI: Compressed Binary Search Tree for Approximate k-NN Searches in Hamming Space. Big Data Research. 2021 Jul 15;25:100223.

Co-Authored Publications:

Jindal A, Gupta P, Jayadeva, Sengupta D. Discovery of rare cells from voluminous single cell expression data. Nature communications. 2018 Nov 9;9(1):1-9.

Gupta P, Jindal A, Jayadeva, Sengupta D. Linear time identification of local and global outliers. Neurocomputing. 2021 Mar 14;429:141-50.

Jindal A, Gupta P, Sengupta D., Jayadeva Enhash: A Fast Streaming Algorithm for Concept Drift Detection. ESANN 2021 proceedings. Online event, 6-8 October 2021, i6doc.com publ., ISBN 978287587082-7.