LEARNING ON LARGE DATASETS USING BIT-STRING TREES

PRASHANT GUPTA

DEPARTMENT OF ELECTRICAL ENGINEERING
INDIAN INSTITUTE OF TECHNOLOGY DELHI

April 2023

LEARNING ON LARGE DATASETS USING BIT-STRING TREES

by

PRASHANT GUPTA

DEPARTMENT OF ELECTRICAL ENGINEERING

Submitted

in fulfillment of the requirements of the degree of Doctor of Philosophy

to the

INDIAN INSTITUTE OF TECHNOLOGY DELHI

April 2023

Certificate

This is to certify that the thesis entitled **"Learning On Large Datasets Using Bit-String Trees"**, being submitted by **Prashant Gupta** for the award of the degree of **Doctor of Philosophy** to the Department of Electrical Engineering, Indian Institute of Technology Delhi, is a record of bonafide work done by him under my supervision and guidance. The matter embodied in this thesis has not been submitted to any other University or Institute for the award of any other degree or diploma.

Dr. Jayadeva

Professor

Department of Electrical Engineering,

Indian Institute of Technology Delhi,

Hauz Khas, New Delhi - 110016,

India.

Dr. Vibhor Kumar

Associate Professor

Department of Computational Biology,

Indraprashta Institute of Information Technology,

Okhla Phase III, Delhi - 110020,

India.

Dr. Debarka Sengupta

Associate Professor

Department of Computer Science & Engineering,

Department of Computational Biology,

Head, Center for Artificial Intelligence,

Indraprashta Institute of Information Technology,

Okhla Phase III, Delhi - 110020,

India.

(Adj.) Associate Professor

Institute of Health & Biomedical Innovation,

QUT, Australia

Acknowledgments

I would like to thank my supervisors Prof. Jayadeva (IITD), Assoc. Prof. Debarka Sengupta (IIITD), Assoc. Prof. Vibhor Kumar (IIITD), Research Committee Members Prof. Indra Narayan Kar (IITD), Prof. Shouri Chatterjee (IITD), and Asst. Prof. Vivekanandan Perumal (IITD); faculty members with whom I worked - Retd. Prof. Suresh Chandra (IITD), Assoc. Prof. Gaurav Ahuza (IIITD), Retd. Prof. Munishwar Nath Gupta (IITD), and Prof. Rajeev Narang (AIIMS); friends and colleagues Aashi Jindal, Dr. Sumit Soman, Dr. Udit Kumar, Aashish Rajiv, Dr. Mayank Sharma, Dr. Shruti Sharma, Dr. Himanshu Pant, Devesh Bajpai, Ritika Bajpai, Sanjay Pandey; department staff members Rakesh Kumar, Yatindra Mani, Mukesh, and Ritwik Pahari; my parents, my parents-in-law, my wife Aashi Jindal and other family members Priyanka Gupta, Sahil Jindal, and Khushboo Mehrotra for supporting me through this journey.

I would like to extend my special thanks to my wife Aashi Jindal for being the part of both journeys, academic and life, and supporting me through the ups of downs.

(Prashant Gupta)

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.

सार

समानता को संरक्षित रखने वाली हैशिंग प्रक्रिया का निकटतम-पडोसी को खोजने में व्यापक अनुप्रयोग मिलता है। समानता-संरक्षण हैशिंग प्रक्रियाओं का अधिकतम उपयोग विस्तार-विभाजन आधारित प्रणालियों में होता है। कई विस्तार-विभाजन आधारित हैशिंग प्रणालियाँ, हैश संकेतावली के रूप में अंश संकेतावली उत्पन्न करती है। हालांकि द्विआधारी खोज वृक्ष (BST) का उपयोग अंश संकेतावलिओं को संगृहित करने के लिए किया जा सकता है, लेकिन इस प्रकार के वृक्षों का आकार संकेतावलिओं की लंबाई के साथ तेजी से बढ़ता है। व्यव्हार में ऐसे वृक्ष अत्यधिक विरल होती है, जिससे निकटतम-पड़ोसिओं को खोजने की क्षति-दर में वृद्धि होती है। BST की विरलता और स्मृति की समस्या के समाधान के लिए हमने उलटे हैश तालिका की सम्पीड़ित द्विआधारी खोज वृक्ष (ComBI) को विकसित किया, जो BST की ज्यामितीय रूप से प्रेरित संपीड़न तकनीक है। ComBI , BST के उल्लेखनीय स्मृति पदचिन्हय के बिना तेज़ और अनुमानित निकटतम-पड़ोसियों की खोज करने में सक्षम है। इस शोध प्रबंध में हम दिखते हैं की हैमिंग विस्तार में ComBI द्वारा अनुमानित रूप से खोजे हुए निकटतम पडोसी, यथार्थ पड़ोसिओं की खोज करने वाली कलन विधिओं के साथ प्रतिस्पर्धी है। व्यापक रूप से उपयोग होने वाले यथार्थ खोज की विधि, बहु सूचकांक हैशिंग (MIH) की तुलना में ComBI ~८ करोड़ प्रतिदर्श वाले आंकडाकोष में ०.९ की औसतन परिशुद्धता विभिन्न लम्बाइयों की संकेतावलिओं पर लगभग ~४ से ~२९६ गुना अधिक तीव्रता से प्राप्त करता है। १ अरब प्रतिदर्ष वाले आंकडाकोष में यह परिशुद्धता MIH की तुलना में लगभग ~४ से ~१९ गुना तीव्रता से प्राप्त हो जाती है। इसके पश्चात, ComBI को एकक कोशिका RNA अनुक्रमण (scRNA-seq) के आंकड़ों द्वारा सामान कोशिकाओं की खोज करने वाले खोज-यन्त्र के रूम में प्रस्तुत किया गया और इसके प्रदर्शन की तुलना अत्याधुनिक scRNA-seq खोज-यन्त्र, Cellfishing.jl से की गयी जो की MIH पे आधारित है। ComBI कई खातों में Cellfishing.jl से बेहतर प्रदर्शन करता है और इस खोज के अभ्यास में लगभग ~२ से ~१३ गुना की तीव्रता भी प्राप्त करता है।

इसके पश्चात हम अपना ध्यान समानता संरक्षण करने वाली हैशिंग विधिओं का उपयोग करके एक वर्गीकर्त्ता को विकसित करने में करते है। इस प्रकार के वर्गिकर्ताओं के निर्माण के लिए हैशिंग कलन विधिओं द्वारा सीखी गयी सरंचना बयेस' वर्गिकर्ताओं (bayes' classifier) के साथ संयुक्त रूप में उपयोग की जा सकती है। इस शोध प्रबंध में हमने तीन मूलभूत विस्तार-विभाजन आधारित हैशिंग कला विधियों का उपयोग करके वर्गिकर्ताओं का निर्माण किया और उनके गुंडों और अवगुडों का विश्लेषण किया है। यह विश्लेषण हमें एक वृक्ष आधारित हैशिंग वर्गीकर्त्ता के निर्माण के लिए प्रेरित करता है। यह शोध प्रबंद वृक्ष आधारित समवेत हैशिंग वर्गीकर्त्ता का उल्लेख करता है जिसे मार्गदर्शित आकस्मिक वर्गीकर्त्ता (GRAF) के नाम से सम्बोधित किया जाता है। GRAF एक परोक्ष विभाजन तकनीक है जो की वृक्ष आधारित वर्गीकरणों का निर्माण सार्वत्रिक विभाजनों के आधार पे करती है। इस प्रकिया में अन्य स्थानीय विभागों का पुनः वर्गीकरण सार्वत्रिक विभाजनों के आधार पे होता है जो की वर्तमान तकनीकों से भिन्न है। हम यह भी दिखते हैं की GRAF, निर्णय वृक्ष (decision tree) और वर्धन कला विधिओं (ensemble algorithms) के मध्यांतर को पूर्ण करता है। प्रयोगों से यह संकेत भी मिलता है कि GRAF सामान्यीकरण त्रुटिओं की सीमा को भी कम करता है। ११५ मापदंड आंकड़ा समूहों के परिणाम दिखते हैं की GRAF अधिकतर आकड़ां समूहों पे उत्तीर्ण परिणाम देता है। इस शोध प्रबंध में अनिरीक्षित GRAF (uGRAF) को भी विकसित किया गया है। मौलिक रूप से GRAF उच्च आंकड़ा जटिल क्षेत्र में अधिक अधिसमतल उत्पन्न करता है और इस घटना को एक प्रतिदर्श को शुद्ध रूप से वर्गीकृत करने के लिए आवश्यक समतलों की संख्या द्वारा दर्शाया जाता है। इस माप का उपयोग महत्वपूर्ण प्रतिदर्श के चयन के लिए किया जा सकता है। इस प्रक्रिया का उपयोग आंकड़ा समूहों के सन्निकटन के लिए भी किया जा सकता है। इस सिद्धांत को स्थापित करने के लिए कृत्रिम और UCI आंकड़ां समूहों के साथ एक व्यापक

अनुभवजन्य मूल्यांकन किया गया है। प्रस्तावित कार्यप्रणाली की तुलना अत्याधुनिक महत्वपूर्ण प्रतिदर्श के चयन करने वाली कला विधियों से की गयी है। इस शोध प्रबंध में शह सदिश प्रतिदर्श (SVs) और GRAF के द्वारा चिन्हित महत्वपूर्ण प्रतिदर्श के मध्य एक समानता भी स्थापित की गयी है।

इसके पश्चात हम दिखाते हैं की प्रतिदर्श के सीखे हुए पड़ोस का उपयोग एक मापनीय तरीके से उसके परिवेश में भ्रम का अनुमान लगाने के लिए किया जाता है। प्रति-प्रतिदर्श वर्गीकरणीयता का अनुमान लगाने के लिए हमने uGRAF और ComBI का उपयोग किया। अनुमानित मूल्यों का एक अनुभवजन्य मूल्यांकन प्रस्तुत किया गया है। हम दिखाते हैं कि कैसे प्रति-प्रतिदर्श वर्गीकरण क्षमता का उपयोग कर्क रोगी की उत्तरजीविता का अनुमान लगाने के लिए किया जा सकता है।

कर्क रोग एक सनजीन का रोग है। सनजीन के उत्परिवर्तन जिनके परिणामस्वरूप कर्क रोग हो सकता है उन्हें विरासत में प्राप्त किया जा सकता है, पर्यावरणीय तत्वों से अधिग्रहित किया जा सकते हैं या यादृच्छिक प्रतिकृति त्रुटियों के परिणामस्वरूप हो सकता है। कार्सिनोजेनेसिस (carcinogenicity) के प्रवेश होने के बाद उत्परिवर्तन फैलता रहता है और कर्क सनजीन में अत्यधिक बदलाव आता है। अधिकांश कर्क रोग सम्बन्धी दैहिक उत्परिवर्तन, जनन रेखा सम्बन्धी उत्परिवर्तन या अन्य गैर-दैहिक उत्परिवत्न से अविभेद्य होते हैं। आज तक केवल एक छोटे से उत्परिवर्तन उपसम्मुचय का विश्लेषण किया गया है। उत्परिवर्तनो का प्रयोगशाला सत्यापन अत्यधिक कठिन और श्रमिक कार्य है। उत्परिवर्तन पार्श्वचित्र में रोगीयोन के मध्य अत्यधिक भिन्नता यह समस्या को और प्रख्यात करती है और रोग विशेष उत्परिवर्तनो को पहचानना और मुश्किल हो जाता है। इसीलिए बहुत सारे हानिकारक उत्परिवर्तन आज भी अज्ञात है। यह शोध प्रबंध एक नयी विधि कोडन स्विचेस का सतत प्रतिनिधित्व (CRCS) का परिचय देता है। यह डीप-लर्निंग (Deep learning) आधारित दृश्टिकोण सनजीन उत्परिवर्तनो के संख्यात्मक सदिश प्रतिनिधित्व का उत्पादन करने में सक्षम बनाता है। यह प्रतिनिधित्व हमें विभिन्न प्रकार के मशीन लर्निंग (machine learning) आधारित कार्यों को करने के सक्षम बनाता है। हम दिखाते हैं कि कैसे CRCS को तीन अलग-अलग तरीकों से इस्तेमाल किया जा सकता है। सबसे पहले हम यह दिखाते हैं की CRCS तकनीक का उपयोग उपयुक्त प्रकृतिस्थ प्रतिदर्श की अनुपस्थिति में दैहिक उत्परिवर्तनो की पहचान करने में किया जा सकता है। दूसरा, प्रस्तुत की गई विधि चालक जीनों को खोजना और उनका अध्ययन करना संभव बनाती है। अंत में, हमने CRCS के साथ एक अनुक्रम वर्गीकारक को जोड़कर उत्परिवर्तनों का एक संख्यात्मक प्रतिनिधित्व बनाया। ये प्रतिनिधियों का उपयोग प्रति प्रतिदर्श के व्यक्तिगत उत्परिवर्तनों के आंकलन में किया जा सकता है। यह आंकलन ब्लैडर यूरोथेलियल कार्सिनोमा (BLCA), हेपेटोसेलुलर कार्सिनोमा (HCC), और ग्लियोब्लास्टोमा मल्टीफॉर्म (GBM) के रोगिओं की जीवन सम्भावना का पूर्वानुमान लगाने में सक्षम था। संछेप में, हम CRCS को व्यक्तिगत कर्क रोग के उत्परिवर्तनों को कार्यात्मक महत्व के विश्लेषण के लिए एक मूल्यवान अभिकलनीय उपकरण के रूप में प्रस्तावित करते हैं।

Contents

List of Figures

List of Tables

List of Algorithms

List of abbreviations

nsr negative sampling rate

ws window size

ADA Adaboost

AP Average Precision

BERT Bidirectional Encoder Representations from Transformers

bi-LSTM Bidirectional Long Short-Tem Memory

BLAC Bidirectional Long Short-Term Memory with Attention & CRCS embeddings

BLCA Bladder Urothelial Carcinoma

BST Binary Search Tree

BSTs Binary Search Trees

cBioPortal cBio Cancer Genomics Portal

cfDNA Cell-free DNA

CGI Cancer Genome Interpreter

CNN Convolutional Neural Network

CO2 Continuously Optimized Oblique

ComBI Compressed BST of Inverted hash tables

COSMIC Catalogue Of Somatic Mutations In Cancer

CPU Central Processing Unit

CRCS Continuous Representation of Codon Switches

CSC Cancer-Stem-Cell

ctDNA Circulating Tumor DNA

DBSCAN Density-Based Spatial Clustering of Applications with Noise

dbSNP Single Nucleotide Polymorphism Database

DNA Deoxyribonucleic Acid

DT Decision Trees

eQTL Expression Quantitative Trait Loci

ESCC Esophageal Squamous Cell Carcinoma

ET Extremely Randomized Trees

EWH Error Weighted Hashing

ExAC Exome Aggregation Consortium

FDR False Discovery Rate

GB Gradient Boosting

GBM Glioblastoma Multiforme

GEPSVM Proximal SVM with Generalized Eigenvalues

gnomAD Genome Aggregation Database

GPU Graphics Processing Unit

GRAF Guided Random Forest

GTO Global Tree Optimization

HBST Hamming Distance Embedding BST

HCC Hepatocellular Carcinoma

HD Hamming Distance

HEPTS Hybrid Extreme Point Tabu Search

HNSC Head and Neck Squamous Cell Carcinoma

HWT Hamming Weight Tree

IBST Inverted-Hash-Table BST

indel Insertions & Deletions

intOgen Integrative Onco Genomics

LDA Linear Discriminant Analysis

LGD Local Gene Duplication

LSB Least Significant Bit

LSH Locality Sensitive Hashing

LSTM Long Short-Tem Memory

LUAD Lung Adenocarcinoma

LUSC Lung Squamous Cell Carcinoma

MIH Multi-Index Hashing

MML Maximum Message Length

MPSVM Multi-Surface Proximal SVM

mRNA Messenger RNA

MRPT Multiple Random Projection Trees

MSB Most Significant Bit

NLP Natural Language Processing

NN Nearest Neighbor

NNs Nearest Neighbors

NPPS Neighborhood Property-Based Pattern Selection

NSCLC Non-Small Cell Lung Cancer

OC1 Oblique Classifier 1

OncoKB Oncology Knowledge Base

OOB Out of Bag

OT Oblique Tree

OTs Oblique decision trees

PCA Principal Component Analysis

PolyPhen2 Polymorphism Phenotyping v2

PR Precision-Recall

RAM Random Access Memory

RF Random Forest

ROI Radius of Influence

RP Random Partition

RVFL Random Vector Functional Link Network

scRNA-seq single-cell RNA sequencing

SIFT Sorting Intolerant From Tolerant

SKCM Skin Cutaneous Melanoma

SNV Single Nucleotide Variant

SOM Self-Organizing Maps

SPH Similarity Preserving Hashing

SV Support Vector

SVIS Small Votes Instance Selection

SVM Support Vector Machine

TMB Tumor Mutational Burden

tSNE t-Distributed Stochastic Neighbor Embedding

UCSC University of California, Santa Cruz

UEC Undifferentiated Endometrial Carcinoma

uGRAF Unsupervised GRAF

VCF Variant Call Format

WES Whole Exome Sequencing

WGS Whole Genome Sequencing

XGB Extreme Gradient Boosting (XGBoost)