Compressed Learning

University of Technology, Sydney

Tianyi Zhou

Faculty of Engineering and Information Technology

University of Technology, Sydney

A thesis submitted for the degree of

Doctor of Philosophy

31 August 2013

To my loving parents
Yongxi Zhou and *Jingping Yao*

CERTIFICATE OF ORIGINAL AUTHORSHIP

I certify that the work in this thesis has not previously been submitted for a degree nor has it been submitted as part of requirements for a degree except as fully acknowledged within the text.

I also certify that the thesis has been written by me. Any help that I have received in my research work and the preparation of the thesis itself has been acknowledged. In addition, I certify that all information sources and literature used are indicated in the thesis.

Student: Tianyi ZHOU
Date: 31/08/2013

Acknowledgements

I benefited and learned a lot from my advisor, several professors, my friends, my colleagues and my family during PhD study in University of Technology, Sydney and Nanyang Technological University. I would like to take this good opportunity to appreciate their significant helps to me.

The first person I would like to express my appreciation and gratitude to is Professor Dacheng Tao. I am so lucky to have Prof. Tao as my professional mentor and academic advisor. I benefited significantly from various detailed discussions with him. Discussing a problem with him has been always a humbling but eye-opening experience, and he always gives me sufficient freedom to think and explore. His careful criticisms bridging both theory and application have greatly broadened my thought in research. His vision, creativeness and enthusiasm in solving challenging problems has greatly encouraged me and inspired my works. Without his high scientific criterion, endless patience, generous support, and constant guidance, this thesis cannot be accomplished. I also want to thank Prof. Tao as a nice friend, for his invaluable suggestions and experienced instructions on my research career and life.

I also wish to express my appreciation to other Professors who have gave me helpful guidance and encouragement during my PhD study and on conferences: Prof. Chengqi Zhang, Prof. Xindong Wu, Prof. Xingquan Zhu, Prof. Jieping Ye, Prof. Jerome H. Friedman, Prof. Trevor Hastie, Prof. Dean P. Foster and Prof. Emmanuel Candès. I learned a lot from discussions with them and their attitude for doing high-quality research.

I have been fortunate to work in a group gathering the most brilliant researchers and best friends in the past 5 years: Dr. Wei Bian, Dr. Jun Li, Dr. Bo Geng, Dr. Xinmei Tian, Dr. Zhang Zhang, Dr. Chao Zhang, Dr. Naiyang Guan, Dr. Lusong Li, Dr. Xiaoguang Rui, Dr. Weifeng Liu, Dr, Shengzheng Wang, Yang Mu, Bo Xie, Dongjing Song, Si Si, Yuanyuan Fu, Yong Luo, Yangxi Li, Zhibin Hong, Nannan Wang, Mingming Gong, Lianyang Ma, Maoying Qiao, Xiaoyan Li, Tongliang Liu, Fei Gao, Changxin Ding, Jie Gui, Xinchao Wang, Tianhao Zhang. Especially, I am deeply indebted to Dr Wei Bian, who gave me strong motivation and critical guidance when I have meet difficulties in research. He spent much time to teach me valuable things that I cannot easily learn by myself, especially at the beginning of my PhD study. I learned lots from discussion and collaboration with all group members in statistics, machine learning and optimization. Moreover, I enjoyed the invaluable friendships with them, their kindly support and accompany are always my source of strength and courage in both research and daily life. I own my deepest thanks to all of them!

I am also grateful to all the other friends who made my 5 years at Singapore and Sydney unforgettable: Guodong Long, Jing Jiang, Chunyang Liu, Meng Fang, Shirui Pan, Mingsong Mao, Hongshu Chen, Can Wang, Junfu Yin, Guoxin Su, Dianshuang Wu, Yi Ji, Ming Xie, Xiang Li, Yifan Li, Qian Sun, and my dearest friends Taoyu Lin and Peng Su since my college. They are the ones who have given me support during both joyful and stressful times, to whom I will always be thankful.

Finally, it is my greatest honor to thank my family: my parents, my grandparents, my uncle and auntie. They are always believing in me, keeping encouraging me, giving me indispensable suggestions, and fully supporting all my final decisions. No words could possibly express my deepest gratitude for their endless love, self-sacrifice and unwavering help. To them I dedicate this dissertation.

Abstract

There has been an explosion of data derived from the internet and other digital sources. These data are usually multi-dimensional, massive in volume, frequently incomplete, noisy, and complicated in structure. These "big data" bring new challenges to machine learning (ML), which has historically been designed for small volumes of clearly defined and structured data. In this thesis we propose new methods of "compressed learning", which explore the components and procedures in ML methods that are compressible, in order to improve their robustness, scalability, adaptivity, and performance for big data analysis. We will study novel methodologies that compress different components throughout the learning process, propose more interpretable general compressible structures for big data, and develop effective strategies to leverage these compressible structures to produce highly scalable learning algorithms. We present several new insights into popular learning problems in the context of compressed learning. The theoretical analyses are tested on real data in order to demonstrate the efficacy and efficiency of the methodologies in real-world scenarios.

In particular, we propose "manifold elastic net (MEN)" and "double shrinking (DS)" as two fast frameworks extracting low-dimensional sparse features for dimension reduction and manifold learning. These methods compress the features on both their dimension and cardinality, and significantly improve their interpretation and performance in clustering and classification tasks.

We study how to derive fewer "anchor points" for representing large datasets in their entirety by proposing "divide-and-conquer anchoring", in which the global solution is rapidly found for near-separable

non-negative matrix factorization and completion in a distributed manner. This method represents a compression of the big data itself, rather than features, and the extracted anchors define the structure of the data.

Two fast low-rank approximation methods, "bilateral random projections (BRP)" of fast computer closed-form and "greedy bilateral sketch (GreBske)", are proposed based on random projection and greedy augmenting update rules. They can be broadly applied to learning procedures that requires updates of a low-rank matrix variable and result in significant acceleration in performance.

We study how to compress noisy data for learning by decomposing it into the sum mixture of low-rank part and sparse part. "GO decomposition (GoDec)" and the "greedy bilateral (GreB)" paradigm are proposed as two efficient approaches to this problem based on randomized and greedy strategies, respectively. Modifications of these two schemes result in novel models and extremely fast algorithms for matrix completion that aim to recover a low-rank matrix from a small number of its entries. In addition, we extend the GoDec problem in order to unmix more than two incoherent structures that are more complicated and expressive than low-rank or sparse matrices. The three proposed variants are not only novel and effective algorithms for motion segmentation in computer vision, multi-label learning, and scoring-function learning in recommendation systems, but also reveal new theoretical insights into these problems.

Finally, a compressed learning method termed "compressed labeling (CL) on distilled label sets (DL)" is proposed for solving the three core problems in multi-label learning, namely high-dimensional labels, label correlation modeling, and sample imbalance for each label. By compressing the labels and the number of classifiers in multi-label learning, CL can generate an effective and efficient training algorithm from any single-label classifier.

Contents

List of Figures

List of Tables