Large-scale Machine Learning Algorithms for Big Data

A DISSERTATION PRESENTED
BY
YAN YAN
TO
CENTRE FOR ARTIFICIAL INTELLIGENCE
FACULTY OF ENGINEERING AND INFORMATION TECHNOLOGY

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY
IN THE SUBJECT OF
COMPUTER SCIENCE

UNIVERSITY OF TECHNOLOGY SYDNEY
ULTIMO, NEW SOUTH WALES
MAY 2018

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.

Signature of Student: Production Note:
Signature removed prior to publication.

Date: 12 Aug 2018

ALL PUBLICATIONS DURING THE CANDIDATURE

Refereed Journal Publications

1. Yan Yan, Feiping Nie, Wen Li, Chenqiang Gao, Yi Yang, Dong Xu. "Image Classification by Cross-Media Active Learning with Privileged Information." *IEEE Transactions on Multimedia* 18, no. 12 (2016): 2494-2502.

2. Yahong Han, Yi Yang, Yan Yan, Zhigang Ma, Nicu Sebe and Xiaofang Zhou. "Semi-Supervised Feature Selection via Spline Regression for Video Semantic Recognition." *IEEE Transactions on Neural Networks and Learning Systems* 26, no. 2 (2015): 252-264.

3. Xingzhong Du, Yan Yan, Pingbo Pan, Guodong Long and Lei Zhao. "Multiple Graph Unsupervised Feature Selection." *Signal Processing* 120 (2016): 754-760.

4. Yan Yan, Gaowen Liu, Sen Wang, Jian Zhang and Kai Zheng. "Graph-Based Clustering and Ranking for Diversified Image Search." *Multimedia Systems* 23, no. 1 (2017): 41-52.

Refereed Conference Publications

1. Xuanyi Dong, Yan Yan, Wanli Ouyang and Yi Yang. Style Aggregated Network for Facial Landmark Detection. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* 2018.

2. Yu Wu, Yutian Lin, Xuanyi Dong, Yan Yan, Wanli Ouyang and Yi Yang. Exploit the Unknown Gradually: One-Shot Video-Based Person Re-Identification by Stepwise Learning. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* 2018.

3. Yan Yan, Tianbao Yang, Yi Yang, Jianhui Chen. A Framework of Online Learning with Imbalanced Streaming Data. In *Thirty-First AAAI Conference on Artificial Intelligence (AAAI)* 2017.

4. Yi Xu*, Yan Yan*, Qihang Lin, Tianbao Yang. Homotopy Smoothing for Non-Smooth Problems with Lower Complexity than O(1/ε). In *Neural Information Processing Systems (NIPS)* 2016.

5. Yan Yan, Zhongwen Xu, Ivor W. Tsang, Guodong Long, Yi Yang. Robust Semi-supervised Learning through Label Aggregation. In *Thirtieth Conference on Artificial Intelligence (AAAI)* 2016.

6. Mingkui Tan, Yan Yan, Li Wang, Anton Van Den Hengel, Ivor W. Tsang, Qinfeng (Javen) Shi. Learning Sparse Confidence-Weighted classifier on Very High Dimensional Data. In *Thirtieth Conference on Artificial Intelligence (AAAI)* 2016.

7. Yan Yan, Mingkui Tan, Ivor W. Tsang, Yi Yang, Chengqi Zhang and Qinfeng (Javen) Shi. Scalable Maximum Margin Matrix Factorization by Active Riemannian Subspace Search. In *International Joint Conference on Artificial Intelligence (IJCAI)* 2015, 3988-3994.

8. Joey Tianyi Zhou, Sinno Jialin Pan, Ivor W. Tsang and Yan Yan. Hybrid Heterogeneous Transfer Learning through Deep Learning. In *Twenty-Eighth Conference on Artificial Intelligence (AAAI)* 2014, 2213-2219.

Thesis advisor: Professor Yi Yang Yan Yan

Large-scale Machine Learning Algorithms for Big Data

Abstract

Machine learning is a research area in artificial intelligence which aims to learn a model from data. On one hand, the target is to learn a model yielding superior performance. On the other hand, as the rapid increase of the size of the collected data, there emerges a demand for machine learning algorithms to deal with large-scale problems.

Recent years have witnessed a sharp increase of the scale of the collected data. Taking recommender systems as an example, the Yahoo Music dataset includes more than 262 million ratings. In image classification, Imagenet contains more than 100 million images from the Internet. Such a large scale brings a great challenge to machine learning algorithms: how could the machine learning algorithms achieve satisfactory performance with less computational cost? In this dissertation, I mainly focus on several specific machine learning tasks and their scalability issues in either computation or storage aspects.

Computational cost plays a crucial role in machine learning algorithms. For instance, iteration complexity is a commonly-used theoretical metric to evaluate how fast an optimization algorithm converges. An example is the full singular value decompositions (SVDs) in the nuclear norm minimization for low-rank matrix completion. Its computational complexity can be $O(n^3)$ where n is the size of the matrix. It would be computationally unfordable when n scales up.

Memory cost is also a typical concern in machine learning. Recently deep neural networks have captured much attention and been successfully applied to a variety of applications. These deep models are known to be hungry for data, so training them usually requires a large number of training samples. When the entire training set cannot be loaded into the memory simultane-

ously, online (stochastic) learning can be applied. In such a memory-restricted scenario, both theoretical analysis and empirical investigation are expected.

Targeting on the above two aspects in large-scale machine learning tasks, in this dissertation, I investigate a variety of machine learning tasks and analyze their specific characteristics. Specifically, I mainly focus on four tasks, i.e., matrix factorization for ordinal ratings, semi-supervised learning, active learning for image classification, online learning for imbalanced streaming data. For the first three tasks, I analyze the specific characteristics of the underlying problems and design new algorithm to optimize the objective. Theoretical verification such as computational complexity is provided. For the last task, I propose an online learning algorithm to deal with imbalanced problems under the strict memory constraint.

Contents

Listing of figures

This thesis is dedicated to my beloved parents and grandparents.

Acknowledgments

First and foremost, I would like to thank my supervisor Professor Yi Yang. I feel extremely fortunate to work with Professor Yi Yang and have learned and progressed so much since the first day when I arrived in Australia. He guided me to investigate a variety of exciting research projects and always encouraged me to look for my own research interests. More importantly, he made me think critically and independently, about research, career and even life. Without this great advice and help during the recent years, it is not possible for me to imagine where I would be and what I would be doing right now. I am sure that I will keep progress under his deep and valuable guidance in the future.

I would like to thank Professor Ivor W. Tsang, Professor Mingkui Tan, Professor Tianbao Yang, Professor Feiping Nie, Professor Dong Xu and Dr Wen Li. I enjoyed the discussions with them. They were always patient to answer my questions. I also learned very much from them, and gradually gained the domain knowledge of many attractive research areas. Their help and advice guided me to find much more interests during my research. The guidance made me increasingly more enthusiastic and eager to contribute more time and energy to my future research.

I want to thank my colleagues and friends in University of Technology Sydney and the University of Iowa. Especially, I would like to thank Zhongwen Xu, Xiaojun Chang, Linchao Zhu, Pingbo Pan, Bo Han, Guoliang Kang, Liang Zheng, Yanbin Liu, Xuanyi Dong, Hu Zhang, Zhun Zhong, Yu Wu, Zhedong Zheng, Yi Xu, Zhe Li, Mingrui Liu, Zaiyi Chen, etc. I was for-

tunate to learn much from them and participate in the wonderful seminars with them. I acquire much knowledges and skills from them, not only the fields that I am working on, but also the areas that I have never touched.