Bandit Learning for Sequential Decision Making

A practical way to address the trade-off between exploration and exploitation

Meng Fang

Faculty of Engineering and Information Technology

University of Technology, Sydney

This dissertation is submitted for the degree of

Doctor of Philosophy

October 2015

To my loving parents.

Declaration

I hereby declare 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 declare 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.

Meng Fang
October 2015

Acknowledgements

There are so many people to thank for helping me during my PhD candidate. So many have made my study in Sydney a lot easier and happier than I thought it was going to be. I wish to express appreciation to all of them.

First, I would like to thank Professor Dacheng Tao for his guidance, encouragement, and patience. Thank you so much for encouraging me to look at research and my work in different ways and for opening my mind. I am so lucky to have Professor Dacheng Tao as my adviser. His great support was essential to my success here.

I would like to thank Professor Xingquan Zhu and Scientist Jie Yin for helping my research and taking time to talk with me on many occasions. I would like to thank Professor Shuliang Wang for introducing Artifical Intelligence research to me.

I would like to thank the members of my faculty: Prof. Chengqi Zhang, Dr. Bin Li, Dr. Lin Chen, Dr. Lu Qin. I learned a lot from discussion with them.

I have been fortunate to work in a group gathering the most brilliant researchers and best friends in the past 4 years: Dr. Wei Bian, Tianyi Zhou, Zhibin Hong, Maoying Qiao, Tongliang Liu, Mingming Gong, Nannan Wang, Associate Professor Weifeng Liu, Associate Professor Bo Du, Ruxin Wang, Qiang Li, Changxin Ding, Zhe Xu, Shaoli Wang, Shujuan Hou, Chang Xu, Chen Gong, Baosheng Yu, Yali Du.

I am also grateful to all the other friends who made my four years at Sydney unforgettable: Chunyang Liu, Bozhong Liu, Shirui Pan, Mingsong Mao, Hongshu Chen, Guodong Long, Jing Jiang, Yifan Fu, Lianhua Chi, Jia Wu, Dianshuang Wu, Yin Song, Can Wang, and my friend Dong Fang, Junhan Gao and Lanfeng Wen since middle school. I would like to especially thank Zhaofeng Su, Allan Yin, Nancy Nan, Hong Man, Hehua Chi, Shaoyuan Li, Xiaoxi Hu, Yinan Li, Wenlin Chen, Shengqi Yang. 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 dearest parents. 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 the dissertation.

Abstract

The sequential decision making is to actively acquire information and then make decisions in large uncertain options, such as recommendation systems and the Internet. The sequential decision becomes challenging since the feedback is often partially observed. In this thesis we propose new algorithms of "bandit learning", whose basic idea is to address the fundamental trade-off between exploration and exploitation in sequence. The goal of bandit learning algorithms is to maximize some objective when making decision. We study several novel methodologies for different scenarios, such as social networks, multi-view, multi-task, repeated labeling and active learning. We formalize these adaptive problems as sequential decision making for different real applications. We present several new insights into these popular problems from the perspective of bandit. We address the trade-off between exploration and exploitation using a bandit framework.

In particular, we introduce "networked bandits" to model the multi-armed bandits with correlations, which exist in social networks. The "networked bandits" is a new bandit model that considers a set of interrelated arms varying over time and selecting an arm invokes the other arms. The objective is still to obtain the best cumulative payoffs. We propose a method that considers both the arm and its relationships between arms. The proposed method selects an arm according to the integrated confidence sets constructed from historical data.

We study the problem of view selection in stream-based multi-view learning, where each view is obtained from a feature generator or source and is embedded in a reproducing kernel Hilbert space (RKHS). We propose an algorithm that selects a near-optimal subset of m views of n views and then makes the prediction based on the subset. To address this problem, we define the multi-view simple regret and study an upper bound of the expected regret for our algorithm. The proposed algorithm relies on the Rademacher complexity of the co-regularized kernel classes.

We address an active learning scenario in the multi-task learning problem. Considering that labeling effective instances across different tasks may improve the generalization error of all tasks, we propose a new active multi-task learning algorithm based on the multi-armed bandits for effectively selecting instances. The proposed algorithm can balance the trade-off

between exploration and exploitation by considering both the risk of multi-task learner and the corresponding confidence bounds.

We study a popular annotation problem in crowdsourcing systems: repeated labeling. We introduce a new framework that actively selects the labeling tasks when facing a large number of labeling tasks. The objective is to identify the best labeling tasks from these noisy labeling tasks. We formalize the selection of repeated labeling tasks as a bandit framework. We consider a labeling task as an arm and the quality of a labeling task as the payoff. We introduce the definition of ε-optimal labeling task and use it to identify the optimal labeling task. Taking the expected labeling quality into account, we provide a simple repeated labeling strategy. We then extend this to address how to identify the best m labeling tasks, and in doing so propose the best m labeling algorithm by indexing the labeling tasks using the expected labeling quality.

We study active learning in a new perspective of active learning. We build the bridge between the active learning and multi-armed bandits. Active learning aims to learn a classifier by actively acquiring the data points, whose labels are hidden initially and incur querying cost. The multi-armed bandit problem is a framework that can adapt the decision in sequence based on rewards that have been observed so far. Inspired by the multi-armed bandits, we consider active learning so as to identify the best hypothesis in an optimal candidate set of hypotheses by involving querying the labels of points as few as possible. Our algorithms are proposed to maintain the candidate set of hypotheses using the error or the corresponding general lower and upper error bounds to help select or eliminate hypotheses. To maintain the candidate set of hypotheses, in the realizable PAC setting, we directly use the error. In the agnostic setting, we use the lower and upper error bounds of the hypotheses. To label the data points, we use the uncertainty strategy based on the candidate set of hypotheses.

Table of contents

List of figures

List of tables