

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 Artificial 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	xv
List of tables	xvii
1 Introduction	1
1.1 Multi-armed bandits	1
1.1.1 Stochastic multi-armed bandit	1
1.2 Networked bandits	4
1.3 Multi-view bandits	5
1.4 Multi-task	6
1.5 Repeated labeling	7
1.6 Active learning	7
1.7 Summary of contributions	8
1.8 Publications	9
2 Networked bandits	11
2.1 Introduction	11
2.2 Related work	14
2.3 Networked bandits	15
2.4 Algorithm	17
2.5 Regret analysis	21
2.6 Practical issues	25
2.6.1 Dynamic network	25
2.6.2 Static network	26
2.6.3 Neighborhood or group	26
2.7 Experiments	26
2.7.1 Illustrative example	26
2.7.2 Baselines and performance metric	27

2.7.3	Simulation experiments	29
2.7.4	Real-world datasets experiments	31
2.8	Conclusion and future work	33
3	Multi-view bandits	35
3.1	Introduction	35
3.2	Related work	39
3.3	Multi-view bandits	42
3.4	CoRLSUB	43
3.4.1	View subset calculation	44
3.5	Regret analysis of CoRLSUB	48
3.6	Experiments	50
3.6.1	Toy example	51
3.6.2	The robot navigation example	52
3.6.3	Public datasets	53
3.6.4	Stream-based multi-view learning	56
3.7	Proofs	57
3.7.1	Proof of Lemma 3.1	57
3.7.2	Proof of Lemma 3.2	58
3.7.3	Proof of Theorem 3.1	59
3.7.4	Proof of Proposition 3.1	60
3.8	Conclusion	60
4	Active multi-task learning via bandits	63
4.1	Introduction	63
4.2	Related work	65
4.3	Problem definition	66
4.4	Algorithm	68
4.4.1	Confidence bounds	70
4.4.2	Active multi-task learning via bandits	72
4.5	Analysis	72
4.6	Experiments	74
4.6.1	Synthetic data	75
4.6.2	Restaurant & consumer data	75
4.6.3	Dermatology data	76
4.6.4	School data	77
4.7	Conclusion	78

5	Selective repeated labeling via bandits	79
5.1	Introduction	79
5.2	Related work	82
5.3	General framework	83
5.4	Algorithm	84
5.4.1	Repeated labeling strategies	85
5.4.2	Selective repeated labeling strategies	89
5.5	Experiments	92
5.5.1	Data sets	92
5.5.2	Labeling strategies	93
5.5.3	Integration methods	93
5.5.4	Results of the selective repeated labeling strategies	93
5.5.5	Comparison between the selective repeated labeling and the single labeling	95
5.5.6	Comparison between the Best m Labeling and the Improved Best m Labeling	96
5.5.7	Study on the size of selected labeling tasks	97
5.6	Conclusion	97
6	Active learning via bandits	99
6.1	Introduction	99
6.2	Related work	101
6.3	Methodology	103
6.3.1	Realizable PAC setting	104
6.3.2	Agnostic setting	105
6.4	Theoretical analysis	107
6.5	Experiments	110
6.5.1	Experimental results of realizable setting	110
6.5.2	Experimental results of agnostic setting	111
6.6	Proofs	112
6.6.1	Proof of Theorem 6.1	112
6.6.2	Proof of Theorem 6.2	114
6.6.3	Proof of Theorem 6.3	115
6.6.4	Proof of Theorem 6.4	116
6.6.5	Proof of Theorem 6.6	116
6.6.6	Proof of Theorem 6.7	117
6.7	Conclusion	119

7 Conclusion	121
References	125

List of figures

2.1	An overview of networked bandits at different rounds. The network is changing over time. An arm (user) can invoke other arms (relations) and has different relations at different rounds. Given the contextual information, the arm is chosen by the decision algorithm for getting multiple payoffs (feedback). The algorithm updates the selection strategy after collecting new payoff information.	12
2.2	An example of the upper bound B in 10-arm networked bandits when $t = 120$. Bar denotes the payoff estimation and vertical line denotes the penalty of the estimation.	21
2.3	An example of the regret value in 10-arm networked bandits. The experiments are repeated 100 times and the average regrets are shown. $y = x$ is provided for comparison.	24
2.4	Illustrative synthetic example of exploration-exploitation trade-off. Bottom, arms with networked topology. Second row: the upper bound B for each arm computed using NetBandits. Third row: the expected estimation v , where bar denotes the estimation and vertical line denotes the penalty of estimation. Fourth row: the real payoff of each arm.	28
2.5	The average payoff at each round in dynamic networks.	30
2.6	The cumulative payoff at each round in dynamic networks.	31
2.7	The average payoff and cumulative payoff for two real-world datasets. . . .	32
3.1	An example of the application of SMVL to the automatic navigation control of a robot.	36
3.2	Using bandit framework to model stream-based multi-view learning.	37
3.3	A toy dataset with different views.	51
3.4	Performance comparison on the toy example.	52
3.5	Performance comparison on robot motion example.	53

3.6	Example views selected by different strategies in the automatic navigation control of a robot.	54
3.7	Performance comparison on (a) G50C and (b) PCMAC.	55
3.8	A comparison of the multi-view bandit strategy with other strategies on Caltech 256.	55
3.9	A comparison of the multi-views bandit strategy with other strategies on VOC 2006.	55
3.10	A comparison of the multi-view bandit strategy with other strategies on ImageNet.	56
4.1	Performance comparison on the synthetic data.	75
4.2	Performance comparison on the Restaurant & consumer data.	76
4.3	Performance comparison on the Dermatology data.	76
4.4	Performance comparison on the School data.	77
5.1	An example of selective repeated labeling. There are a large number of labeling tasks, where each task corresponds to multiple repeated labels and an integrated label (using a majority voting/average ratings). Our goal is to design a selective repeated labeling strategy that identifies the best m labeling tasks.	80
5.2	A comparison of test accuracy between the Best m Labeling and the Random on different data sets.	94
5.3	A comparison of test accuracy between the Best m Labeling and the Improved Best m Labeling on different data sets.	95
5.4	The test accuracy of the Best m Labeling and the Random on different data sets.	96
6.1	Labeled data points rate.	111
6.2	Test error rates for the classification experiments.	111
6.3	Labeled data points rate.	112
6.4	The locations of label queries. The x-axis is the unit interval and the y-axis is the rate of numbers in the corresponding interval. The top histogram shows the locations of label requests at the early stage; the bottom histogram is for all label queries.	113
6.5	Test error rates for the classification experiments.	114

List of tables

2.1	Running time results of NetBandits on four synthetic datasets.	30
5.1	The 9 data sets used in the experiments, including the numbers of attributes, and examples in each, and the split into positive and negative examples. . .	93
5.2	The test accuracy of the Best m Labeling and the Single Labeling.	97

