Advanced Topics in Multi-label Learning

Weiwei Liu

Faculty of Engineering and Information Technology

University of Technology Sydney

A thesis submitted for the degree of

Doctor of Philosophy

July, 2017

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: Weiwei Liu
Date: 24/07/2017

I would like to dedicate this thesis to my loving grandparents and wife.

Acknowledgements

I would like to express my deepest gratitude to my supervisor Prof. Ivor W.Tsang for his patient and valuable guidance. I really appreciate that Prof. Ivor W.Tsang provided me an opportunity to do research under his supervision, which means the change point to me and my life. I knew very little about machine learning and research when I follow Prof. Ivor W.Tsang. It was Prof. Ivor W.Tsang who taught me how to do research, how to find interesting ideas, how to develop fancy models and algorithms, how to write technical papers and how to become an independent researcher all from scratch. He gave me too much patience, and was very willing to teach everything he knows to me. I remember that Prof. Ivor W.Tsang even helped me to fix the bugs for my first research project, and rewrite the technical papers. Without his illuminating instructions, insightful inspiration, consistent encouragement, and expert guidance, I would not have published papers on the leading journals or conferences in my research field. Therefore, I feel very lucky to be supervised by Prof. Ivor W.Tsang.

I am also greatly indebted to the Centre for Quantum Computation & Intelligent Systems (QCIS) directed by Prof. Chengqi Zhang. QCIS has supported me to attend many prestigious international conferences, such as AAAI and NIPS. I really thank QCIS for the support. I also want to express my gratitude to all the students in QCIS.

Last but not the least, I also want to express my deepest gratitude to my wife, Xiuwen Gong. She has accompanied me for six years poor life. She never complains, and always gives me too much encouragement and patience. She is very smart. I like to talk with my wife about my problems, and she always gives me inspirations. During these years, we met with many problems. But, I still feel happiness. Without her support and patience, I can not make any achievements, and also can not live a happy life. I feel extremely grateful for my wife's consistently supporting, encouraging and caring for me all of my life!

Abstract

Multi-label learning, in which each instance can belong to multiple labels simultaneously, has significantly attracted the attention of researchers as a result of its wide range of applications, which range from document classification and automatic image annotation to video annotation.

Many multi-label learning models have been developed to capture label dependency. Amongst them, the classifier chain (CC) model is one of the most popular methods due to its simplicity and promising experimental results. However, CC suffers from three important problems: Does the label order affect the performance of CC? Is there any globally optimal classifier chain which can achieve the optimal prediction performance for CC? If yes, how can the globally optimal classifier chain be found? It is non-trivial to answer these problems. Another important branch of methods for capturing label dependency is encoding-decoding paradigm. Based on structural SVMs, maximum margin output coding (MMOC) has become one of the most representative encoding-decoding methods and shown promising results for multi-label classification. Unfortunately, MMOC suffers from two major limitations: 1) Inconsistent performance: D. McAllester has already proved that structural SVMs fail to converge on the optimal decoder even with infinite training data. 2) Prohibitive computational cost: the training of MMOC involves a complex quadratic programming (QP) problem over the combinatorial space, and its computational cost on the data sets with many labels is prohibitive. Therefore, it is non-trivial to break the bottlenecks of MMOC, and develop efficient and consistent algorithms for solving multi-label learning tasks. The prediction of most multi-label learning methods either scales linearly with the number of labels or involves an expensive decoding process, which usually requires solving a combinatorial optimization. Such approaches become unacceptable when tackling thousands of labels, and are impractical for real-world applications, such as document annotation. It is imperative to design an efficient, yet accurate multi-label learning algorithm with the minimum number of predictions. This thesis systematically studies how to efficiently solve aforementioned issues with provable guarantee.

Contents

Contents v

List of Figures viii

List of Tables x

1 Introduction 1
- 1.1 Background . . . 1
- 1.2 Advanced Topics . . . 4
  - 1.2.1 Underlying Problems Behind CC . . . 5
  - 1.2.2 Major Limitations of MMOC . . . 5
  - 1.2.3 Scalability of Prediction . . . 6
- 1.3 Thesis Contributions . . . 6
  - 1.3.1 Underlying Problems Behind CC . . . 6
  - 1.3.2 Major Limitations of MMOC . . . 7
  - 1.3.3 Scalability of Prediction . . . 7
- 1.4 Thesis Outline . . . 8
- 1.5 Publications . . . 8

2 Literature Review 10
- 2.1 Problem Transformation . . . 10
- 2.2 Algorithm Adaptation . . . 12
- 2.3 Classifier Chain . . . 14
- 2.4 Maximum Margin Output Coding . . . 17
- 2.5 Prediction Complexity Categories . . . 19
  - 2.5.1 Huffman Coding . . . 21
  - 2.5.2 Shannon-Fano Coding . . . 22
- 2.6 Evaluation Metric . . . 22

3 On the Optimality of Classifier Chain for Multi-label Classification 25
- 3.1 Motivations . . . 25

3.2 Proposed Model and Generalization Error Analysis 27
3.2.1 Generalized Classifier Chain 27
3.2.2 Generalization Error Analysis 27
3.3 Optimal Classifier Chain Algorithm 32
3.3.1 Dynamic Programming Algorithm 32
3.3.2 Greedy Algorithm 33
3.3.3 Tree-Based Algorithm 33
3.4 Complexity Analysis 34
3.5 Experiment 35
3.5.1 Data Sets and Baselines 35
3.5.2 Prediction Performance 36
3.5.3 Training Time and Testing Time 37
3.5.4 Results of Many Labels 38
3.5.5 Comparisons with Deep Learning Methods 38
3.6 Summary of This Chapter 42

4 Large Margin Metric Learning for Multi-label Classification 44
4.1 Motivations 44
4.2 Large Margin Metric Learning 46
4.2.1 Preliminaries 46
4.2.2 Proposed Formulation 47
4.2.3 Accelerated Proximal Gradient Update 48
4.2.4 Prediction 49
4.2.5 Complexity Analysis 50
4.3 Generalization Error Analysis 51
4.4 Experiment 53
4.4.1 Experimental Setup 53
4.4.1.1 Data Sets 53
4.4.1.2 Baseline Methods 54
4.4.2 Prediction Performance 55
4.4.3 Comparison with DML 56
4.4.4 Training Time and Testing Time 57
4.5 Summary of This Chapter 57

5 Fast Prediction via Multi-Label Coding Tree 59
5.1 Motivations 59
5.2 Coding Tree Framework 62
5.2.1 Label Powerset Prediction 62
5.2.2 Multi-label Prediction 62
5.3 Theoretical Analysis 65
5.3.1 Analysis on Number of Predictions 65

5.3.2 Testing Time Complexity Analysis 67
5.4 Experiment 67
5.4.1 Data Sets and Baselines 67
5.4.2 Prediction Performance 68
5.4.3 Testing Time 69
5.5 Summary of This Chapter 70

6 Conclusion and Future Work **71**
6.1 Conclusion 71
6.2 Future Work 73

A Appendix **77**
A.1 Covering Numbers 77
A.2 Proof of Lemma 2 78
A.3 Proof of Theorem 3 79
A.4 CC-Greedy algorithm 79
A.5 Proof of Theorem 6 81
A.6 Proof of Lemma 3 83

References **85**

List of Figures

1.1 The illustration of image annotation, where an image may have *cloud*, *tree* and *sky* tags. 2
1.2 Some multi-labeled examples from TRECVID data set. T and F represent the positive and negative labels for corresponding concepts respectively. 3
1.3 The organization of this thesis. 8

2.1 An example of multi-label samples with five labels. 11
2.2 The illustration of *copy* transformation. 14
2.3 The illustration of *copy-weight* transformation. 15
2.4 The illustration of *select-max* transformation. 16
2.5 The illustration of *select-min* transformation. 16
2.6 The illustration of *select-random* transformation. 17
2.7 The illustration of *ignore* transformation. 17
2.8 The first data set produced by the BR method. 18
2.9 The second data set produced by the BR method. 18
2.10 The third data set produced by the BR method. 19
2.11 The fourth data set produced by the BR method. 19
2.12 The fifth data set produced by the BR method. 20
2.13 The transformed data set using the label powerset method. 20
2.14 An example of obtaining a ranking among labels using label powerset method with probability outputs. 21
2.15 Categorization of representative evaluation metric used in this thesis. . 22
2.16 Categorization of representative algorithm adaptation and problem transformation algorithms reviewed in this Chapter. 24

3.1 The training time of CC-DP, CC-Greedy and other baselines on various data sets. yahoo_art, eurlex_sm_10 and eurlex_ed_10 are abbreviated to ART, SM and ED, respectively. 39

3.2 The testing time of CC-DP, CC-Greedy and other baselines on various data sets. yahoo_art, eurlex_sm_10 and eurlex_ed_10 are abbreviated to ART, SM and ED, respectively. 39
3.3 The training and testing time of BR, CC, ECC, Tree-Greedy and Tree-DP on eurlex_sm and eurlex_ed data sets. 41

4.1 The training time of LM-kNN and other baseline methods on all data sets. 56
4.2 The testing time of LM-kNN and other baseline methods on all data sets. 57

5.1 **Top**: Frequency of each label on the delicious and Eur-Lex(ed) data sets. **middle**: Frequency of each label powerset on the delicious and Eur-Lex(ed) data sets. **bottom**: Frequency of samples with the specific number of labels on the delicious and Eur-Lex(ed) data sets. 60
5.2 Schematic illustration of HCT. 63
5.3 Schematic illustration of SFCT. 65
5.4 Testing time of HCT, SFCT and other baseline methods on all data sets (EUR-Lex is abbreviated to EUR). 69

6.1 The conclusions of this thesis. 74

List of Tables

1.1 The applications of multi-label learning. 5

3.1 Time complexity comparisons among CC-Greedy, CC-DP and other baselines. 35
3.2 Data sets used in the experiments of Chapter 3. 36
3.3 The Example-F1 results of CC-Greedy, CC-DP and other baselines on the various data sets (mean ± standard deviation). The best results are in bold. Numbers in square brackets indicate the rank. 36
3.4 The Macro-F1 results of CC-Greedy, CC-DP and other baselines on the various data sets (mean ± standard deviation). The best results are in bold. Numbers in square brackets indicate the rank. 37
3.5 The Micro-F1 results of CC-Greedy, CC-DP and other baselines on the various data sets (mean ± standard deviation). The best results are in bold. Numbers in square brackets indicate the rank. 38
3.6 The Example-F1 results on eurlex_sm and eurlex_ed data sets (mean ± standard deviation). The best results are in bold. Numbers in square brackets indicate the rank. "-" denotes the training time is more than one week. 40
3.7 The Macro-F1 results on eurlex_sm and eurlex_ed data sets (mean ± standard deviation). The best results are in bold. Numbers in square brackets indicate the rank. "-" denotes the training time is more than one week. 40
3.8 The Micro-F1 results on eurlex_sm and eurlex_ed data sets (mean ± standard deviation). The best results are in bold. Numbers in square brackets indicate the rank. "-" denotes the training time is more than one week. 41
3.9 Testing error rate (in %) of VGG, ResNet-34, ADIOS and CCMC-FG on the ILSVRC2012 data set. 42

4.1 Time complexity comparisons between LM-kNN and other baselines. 50
4.2 Data sets used in the experiments of Chapter 4. 53

4.3 The Hamming Loss results of LM-kNN and other baselines on the various data sets (mean $\pm$ standard deviation). The best results are in bold. 55
4.4 The Micro-F1 results of LM-kNN and other baselines on the various data sets (mean $\pm$ standard deviation). The best results are in bold. . . 55
4.5 The Example-F1 results of LM-kNN and other baselines on the various data sets (mean $\pm$ standard deviation). The best results are in bold. . . 55
4.6 Comparison between DML and LM-kNN in terms of Micro-F1 and Example-F1 (mean $\pm$ standard deviation). The best results are in bold. 56

5.1 Testing time complexity comparisons among HCT, SFCT and other baselines. $\Upsilon, \Psi, \iota, \mathfrak{D}$: # clusters, the average # instances in each cluster, # learners and the dimension of the embedding space used in SLEEC. 67
5.2 Data sets used in the experiments of Chapter 5. 68
5.3 The Example-F1 Results of HCT, SFCT and other baselines on the various data sets (mean $\pm$ standard deviation). The best results are in bold. Numbers in square brackets indicate the rank. ”-” indicates that we can not get the results within one week. 69