

University of Technology Sydney
Faculty of Engineering and Information Technology
The Hong Kong Polytechnic University
Department of Computing

Learning Sparse Graphical Models for Data Restoration and Multi-Label Classification

Qiang Li

A thesis submitted in partial fulfilment of the requirements for
the degree of

Doctor of Philosophy

April 2017

CERTIFICATE OF ORIGINALITY

I hereby declare that this thesis is my own work and that, to the best of my knowledge and belief, it reproduces no material previously published or written, nor material that has been accepted for the award of any other degree or diploma, except where due acknowledgement has been made in the text.

_____ (Signed)

_____ Qiang Li (Name of student)

To my loving parents.

Abstract

Sparse probabilistic graphical models play an important role in structured prediction when the dependency structure is unknown. By inducing sparsity over edge parameters, a typical sparse graphical model can combine structure learning and parameter estimation under a unified optimization framework. In this thesis, we propose three specific sparse graphical models accompanied by their applications in data restoration and multi-label classification respectively.

For the data restoration task, we propose random mixed field (RMF) model to explore mixed-attribute correlations among data. The RMF model is capable of handling mixed-attribute data denoising and imputation simultaneously. Meanwhile, RMF employs a structured mean-field variational approach to decouple continuous-discrete interactions to achieve approximate inference. The effectiveness of this model is evaluated on both synthetic and real-world data.

For the multi-label classification task, we propose correlated logistic model (CorrLog) and conditional graphical lasso (CGL), to learn conditional label correlations. (1) The CorrLog model characterizes pairwise label correlations via scalar parameters, thus effects in an explicit (or direct) fashion. More specifically, CorrLog extends conventional logistic regression by jointly modelling label correlations. In addition, elastic-net regularization is employed to induce sparsity over the scalar parameters that define label correlations. CorrLog can be efficiently learned by regularized maximum pseudo likelihood estimation which

enjoys a satisfying generalization bound. Besides, message passing algorithm is applied to solve the multi-label prediction problem. (2) The CGL model further leverages features in modelling pairwise label correlations in terms of parametric functions of the input features, which effects in an implicit (or indirect) fashion. In general, CGL provides a unified Bayesian framework for structure and parameter learning conditioned on input features. We formulate the multi-label prediction as CGL inference problem, which is solved by a mean field variational approach. Meanwhile, CGL learning is efficient after applying the maximum a posterior (MAP) methodology and solved by a proximal gradient procedure. The effectiveness of CorrLog and CGL are evaluated on several benchmark multi-label classification datasets.

Acknowledgements

I feel so grateful to have Prof. Dacheng Tao and Prof. Jane You as my supervisors in UTS and PolyU respectively. I have learned a lot from their immense and extensive knowledge, meticulous attitude to academic research, and genial personalities. Besides, I am grateful to have Dr. Wei Bian and Dr. Richard Yi Da Xu as my de-facto co-supervisors in UTS. I have also learned a lot from their insightful guidance and elaborate suggestions during the second and third years of my PhD candidature. Actually, without those fruitful discussions, I can hardly achieve the research outcomes which constitute the main content of this thesis.

I would like to thank my friends in UTS, including Maoying Qiao, Mingming Gong, Ruxin Wang, Zhibin Hong, Tongliang Liu, Changxing Ding, Meng Fang, Zhe Xu, Shaoli Huang, Hao Xiong, Xiyu Yu, Guoliang Kang, Liu Liu, Chen Gong, Huan Fu, Baosheng Yu, Zhe Chen, Zijing Chen, Yali Du, Xun Yang, Jiankang Deng, Jing Yang, Dr. Jun Li, Dr. Tianyi Zhou, Dr. Yong Luo, Chang Xu, A/Prof. Shengzheng Wang, A/Prof. Xianhua Zeng, Prof. Bo Du, A/Prof. Shigang Liu, A/Prof. Wankou Yang, A/Prof. Xianye Ben, A/Prof. Tao Lei, Liping Xie, Jun Li, Long Lan, Chunyang Liu, Bozhong Liu, Sujuan Hou, Wuxia Yan, Haishuai Wang, Qin Zhang, Zhaofeng Su, Zhiguo Long, Lianyang Ma, Nannan Wang, Fei Gao, Weilong Hou, Mingjin Zhang, Chao Ma and many others. I learned a lot from the discussions and interactions with them in both academic research and

daily life. In particular, I really enjoyed the relaxed atmosphere of the afternoon meet-ups with Dr. Wei Bian and Maoying Qiao. Thanks for their understanding and tolerance to my tedious complaints on research, and thanks for their help in my preparation of research papers. Another particular thanks goes to Mingming Gong and Ruxin Wang, thanks for their encouragements and help on my PhD application, comments and suggestions on my research, and considerate assistance in daily life.

I would also like to thank my friends in PolyU, including A/Prof. Lefei Zhang, Ruohan Zhao, Siwei Hu, Yanxin Hu, A/Prof. Risheng Liu, Dr. Xianbiao Qi, Hui Li, Runjie Tan, Wengen Li, Ruosong Yang, Wei Lu, Yumeng Guo, Xiao Shen, Sitong Mao, Jiaxin Chen, Minglei Li, Edison Chan, Qiang Zhang, Liang Zhang, Quanyu Dai, Zimu Zheng, Yu Lei, Lei Han, Lei Xue, Zhijian He, Xingye Lu and many others. Those occasional discussions and talks with them made the fourth year of my PhD study a nice journey.

Finally, my special thanks goes to my family, including my parents, my elder brother, my grandmother and also in memory of my grandfather who left us three years ago. Thanks for their endless love, encouragements, support and blessing that helped me make it through the hard times of my study and life.

Table of Contents

Table of Contents	x
List of Figures	xiii
List of Tables	xvi
1 Introduction	1
1.1 Background	1
1.2 Continuous Graphical Models	2
1.3 Discrete Graphical Models	3
1.4 Mixed Graphical Models	5
1.5 Data Restoration	6
1.5.1 Data Denoising Methods	7
1.5.2 Data Imputation Methods	9
1.5.3 Model Induction with Corrupted Data	10
1.6 Multi-Label Classification	12
1.6.1 The View of Label Dependence	12
1.6.2 The View of Learning Strategy	13
1.6.3 Music Annotation and Retrieval	13
1.6.4 Multi-Label Image Classification	14
1.7 Summary of Contributions	15

TABLE OF CONTENTS

2	Random Mixed Field Model for Data Restoration	17
2.1	Introduction	18
2.2	Related Works	20
2.3	Random Mixed Field Model	21
2.4	Algorithms	24
2.4.1	Structured Mean Field	24
2.4.2	Parameter Estimation	27
2.5	Experiments	30
2.5.1	Evaluation on Synthetic Data	30
2.5.2	Evaluation on Real Data	32
2.5.3	Data Denoising	33
2.5.4	Noisy Data Imputation	33
2.5.5	Noisy Data Imputation After Denoising	36
2.6	Summary	45
3	Elastic-Net Correlated Logistic Model for Multi-Label Classification	46
3.1	Introduction	47
3.2	Related Works	48
3.3	Correlated Logistic Model	50
3.3.1	Correlated Logistic Regressions	51
3.3.2	Elastic Net Regularization	52
3.4	Algorithms	53
3.4.1	Approximate Learning via Pseudo Likelihood	54
3.4.2	Joint Prediction by Message Passing	58
3.5	Generalization Analysis	58
3.5.1	The Stability of MPLE	60
3.5.2	Generalization Bound	64
3.5.3	Empirical Evaluation	67

TABLE OF CONTENTS

3.6	Experiments: Music Annotation and Retrieval	68
3.6.1	Experimental Setting	70
3.6.2	Results and Dicussions	71
3.7	Experiments: Multi-Label Image Classification	71
3.7.1	A Warming-Up Qualitative Experiment	73
3.7.2	Quantitative Experimental Setting	74
3.7.3	Quantitative Results and Discussions	75
3.7.4	Complexity Analysis and Execution Time	77
3.8	Summary	78
4	Conditional Graphical Lasso for Multi-Label Classification	81
4.1	Introduction	82
4.2	Related Works	84
4.3	Model Representation	85
4.3.1	Graphical Lasso	86
4.3.2	Conditional Graphical Lasso	86
4.4	Algorithms	88
4.4.1	Approximate Inference	89
4.4.2	Structure and Parameter Learning	91
4.5	Experiments	95
4.5.1	Label Graph Structure of CGL	96
4.5.2	Comparison Methods and Measures	97
4.5.3	Results and Discussion	99
4.6	Summary	101
5	Conclusions	106
5.1	Summary of This Thesis	106
5.2	Future Works	107
	References	109

List of Figures

2.1	An example of random mixed field model. (a) The hidden network is a “mixed-net” consisting of both continuous and discrete nodes. (b) explains all the four types of nodes and five types of edges.	21
2.2	The proposed structured mean field approximation can be regarded as cutting off those mixed-type edges and absorbing the interactions in the form of expected sufficient statistics, i.e., $\mathbb{E}_{q(u_s)}[u_s]$ and $\mathbb{E}_{q(v_j)}[\rho_{sj}(v_j)]$, respectively. Such a posterior approximation will result in two separate subgraphs, which are much easier to handle. In addition, it is required to alternately update each of the two subgraphs’ joint distributions until convergence.	24
2.3	The mixed-net graph used in our simulation contains 15 continuous (HSV-colored) and 10 discrete (grey-colored) nodes. The nodes are colored according to attribute values of a representative example. The three types of edges (continuous-continuous in red, discrete-discrete in black and continuous-discrete in light grey) are randomly chosen from all possible edges.	30
2.4	KNN (left plot) and SVM (right plot) classification accuracies of noisy (black) and denoised (light grey) data under different levels of random noise (the noise strength τ ranges from 0.1 to 0.5). Each bar represents the mean and standard deviation of 10 independent experiments.	31

LIST OF FIGURES

2.5	The degeneration curves of classification accuracy versus number of missing attributes on “Adult” dataset: Imputation + KNN Classifier	37
2.6	The degeneration curves of classification accuracy versus number of missing attributes on “Adult” dataset: Imputation + SVM Classifier	38
2.7	The degeneration curves of classification accuracy versus number of missing attributes on “Credit” dataset: Imputation + KNN Classifier	39
2.8	The degeneration curves of classification accuracy versus number of missing attributes on “Credit” dataset: Imputation + SVM Classifier	40
2.9	The degeneration curves of classification accuracy versus number of missing attributes on “Statlog-Australian” dataset: Imputation + KNN Classifier	41
2.10	The degeneration curves of classification accuracy versus number of missing attributes on “Statlog-Australian” dataset: Imputation + SVM Classifier	42
2.11	The degeneration curves of classification accuracy versus number of missing attributes on “Statlog-German” dataset: Imputation + KNN Classifier	43
2.12	The degeneration curves of classification accuracy versus number of missing attributes on “Statlog-German” dataset: Imputation + SVM Classifier	44
3.1	Empirical evaluation of the generalization bound of CorrLog with different number of labels.	67

LIST OF FIGURES

4.1	Comparison of graphical models between unconditional and conditional graphical Lasso. The templates denotes replica of n training images and labels. $\mathbf{x}^{(l)}$ represents the l -th image and $\mathbf{y}^{(l)}$ denotes its label vector. The parameters $\{\boldsymbol{\nu}, \boldsymbol{\omega}\}$, $\{\boldsymbol{\alpha}, \boldsymbol{\beta}\}$ are shared across training data, and are themselves parameterized by hyperparameters λ_1 and λ_2 . In graphical Lasso, $\boldsymbol{\nu}$ and $\boldsymbol{\omega}$ parameterize unary and pairwise potentials, respectively. In contrast, the parameterization is achieved by considering linear functions of $\mathbf{x}^{(l)}$, i.e., $\boldsymbol{\beta}^T \mathbf{x}^{(l)}$ and $\boldsymbol{\alpha}^T \mathbf{x}^{(l)}$, in conditional graphical Lasso.	82
4.2	Illustration of the CGL label graphs learned from PASCAL07-CNN.	102
4.3	Illustration of the CGL label graphs for test example images in PASCAL07-CNN.	103
4.4	Performance variation of CGL versus the hyperparameter λ_2 on MULANscene.	104
4.5	Performance variation of CGL versus the hyperparameter λ_2 on PASCAL07-CNN.	105

List of Tables

2.1	Datasets Summary. #Train./#Test. Inst. Stands for the Number of Training and Testing Instances Respectively. #Num./#Cat. Attr. Stands for the Number of Numerical and Categorical Attributes Respectively.	33
2.2	Classification Accuracies with/without Data Denoising.	34
2.3	Classification Accuracies with Noisy Data Imputation when $\tau = 0.2$	34
2.4	Classification Accuracies with Noisy Data Imputation when $\tau = 0.4$	35
3.1	Summary of important notations for generalization analysis.	59
3.2	Experimental results for top 97 popular tags. CBA stands for Codeword Bernoulli Average (CBA) [51], GMM for Gaussian Mixture Models [128], DirMix for Dirichlet Mixture model [92].	69
3.3	Experimental results for top 78 popular tags. CBA stands for Codeword Bernoulli Average (CBA) [51], HEM-GMM for hierarchical EM Gaussian Mixture Models [128], HEM-DTM for hierarchical EM Dynamic Texture Model [31].	69
3.4	Datasets summary. #images stands for the number of all images, #features stands for the dimension of the features, and #labels stands for the number of labels.	72
3.5	Learned CorrLog label graph on MITscene using ℓ_2 or elastic net regularization.	73

LIST OF TABLES

3.6	MULANscene performance comparison via 5-fold cross validation. Marker */⊗ indicates whether CorrLog is statistically superior/inferior to the compared method (using paired t-test at 0.05 significance level).	75
3.7	MITscene performance comparison via 5-fold cross validation. Marker */⊗ indicates whether CorrLog is statistically superior/inferior to the compared method (using paired t-test at 0.05 significance level). 76	76
3.8	PASCAL07 performance comparison via 5-fold cross validation. Marker */⊗ indicates whether CorrLog is statistically superior/inferior to the compared method (using paired t-test at 0.05 significance level).	77
3.9	PASCAL12 performance comparison via 5-fold cross validation. Marker */⊗ indicates whether CorrLog is statistically superior/inferior to the compared method (using paired t-test at 0.05 significance level).	77
3.10	Computational complexity analysis. Recall that n stands for the number of train images, D stands for the dimension of the features, and m stands for the number of labels. Note that C is the iteration number of the max-product algorithm in CorrLog, and K is the number of nearest neighbours in MLKNN and IBLR.	79
3.11	Average execution time (in seconds) comparison on MULANscene and MITscene.	79
3.12	Average execution time (in seconds) comparison on PASCAL07 and PASCAL12.	79
4.1	Datasets summary. #images stands for the number of all images, #features stands for the dimension of the features, and #labels stands for the number of labels.	96

LIST OF TABLES

4.2	Multi-label image classification performance comparison on MU- LANscene via 5-fold cross validation	98
4.3	Multi-label image classification performance comparison on PAS- CAL07 via 5-fold cross validation	98
4.4	Multi-label image classification performance on PASCAL12 com- parison via 5-fold cross validation	98