Causal and Causally-inspired Learning

Mingming Gong

Faculty of Engineering and Information Technology

University of Technology Sydney

A thesis submitted for the degree of

Doctor of Philosophy

2017

To my loving parents, wife, and son.

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.

Mingming Gong
28/03/2017

Acknowledgements

I would like to sincerely thank everyone who has helped me to finish my doctoral studies.

First of all, I would like to express my sincere appreciation and deep gratitude to my supervisor **Prof. Dacheng Tao**. He has given me trust and freedom to pursue my research interests, and provided constructive suggestions to help me out of difficulties. I can always benefit and learn a lot from various detailed discussions with him, and be excited and energised by his amazing insight, unlimited patience, generous support, and constant encouragement. I feel very lucky to have had him as my supervisor.

I also wish to express my sincere appreciation to my co-supervisor **Dr. Kun Zhang** who hosted my visit to Max Planck Institute for Intelligent Systems. He has been leading me and my research with his passion, patience, optimism, intelligence, and creativity. His research styles and way of thinking have influenced me very deeply.

I also would like to give special thanks to my excellent collaborators: Prof. Bernhard Schölkopf, Prof. Clark Glymour, Dr. Dominik Janzing, Prof. Yuhong Wang, Prof. Changyin Sun, A/Prof. Di Huang, A/Prof. Junliang Xing, A/Prof. Wankou Yang, Dr. Chaohui Wang, Dr. Tongliang Liu, Dr. Chang Xu, Mr. Philipp Geiger, Mr. Ruxing Wang, Mr. Qiang Li, Mr. Shaoli Huang, Mr. Huan Fu, for their brilliant work and timely support. I have also been fortunate to work and have discussions with many other brilliant researchers: Dr. Nannan Wang, Dr. Fei Gao, Prof. Chen Gong, Dr. Yong Luo, Dr. Lianyang Ma, Dr. Weilong Hou, A/Prof. Bo Du, A/Prof. Shigang Liu, A/Prof.

Xianye ben, and Dr. Qiong Wang. I wish to express appreciation to all of them for their support and kind accompany.

I am so grateful to my labmates and friends in Sydney: Tongliang Liu, Ruxin Wang, Shaoli huang, Qiang Li, Zhibin Hong, Meng Fang, Changxing Ding, Maoying Qiao, Zhe Xu, Long Lan, Wei Bian, Tianyi Zhou, Jun Li, Chunyang Liu, Bozhong Liu, Zhiguo Long, Guodong Long, Jing Jiang, Huan Fu, Baosheng Yu, Zhe Chen, Zijing Chen, Xiyu Yu, Liu Liu, Yali Du, Guoliang Kang, Hao Xiong, Jiayan Qiu, Jiankang Deng, Chaoyue Wang, Dayong Tian, Jiang Bian, Bo Han, Kede Ma, Tianrong Rao, Lingxiang Wu, Sujuan Hou, Xiaoqing Yin, Peicheng Zhou, Tao Zhang, Haifeng Liu, Peng Hao, Haishuai Wang, and Xun Yang. I am especially deeply indebted to Tongliang, who has given me inspirations on the theoretical problems. I would also like to express my sincere thanks to Ruxin, who has helped me a lot in my daily life. All my friends here have provided strong support during both happy and stressful times. I am also grateful to friends I met in Tübingen: Hueihan Jhuang, Chaohui Wang, Zhitang Chen, Zhikun Wang, Tingting Shao, Yanlong Huang, Junyi Zhu, Pengsheng Zheng, Francesco Dinuzzo, and Krikamol Muandet. I had spent a wonderful time with them in Tübingen.

Finally, I would like to express deep-felt gratitude to my family: my parents, my wife, and my son for their endless love, trust, encouragement, and full support throughout my studies and life. I dedicate this thesis to them.

Abstract

A main goal of statistics and machine learning is to discover statistical dependencies between random variables, and these dependencies will be used to perform predictions on future observations. However, many scientific investigations involve causal predictions, the aim of which is to infer how the data generating system should behave under changing conditions, for example, changes induced by external interventions. To perform causal predictions, we need both statistical dependencies as well as causal structures to determine the behaviour of the system. The standard way to identify causal structures is to use randomized controlled experiments. However, conducting these experiments is usually expensive or even impossible in many scenarios. As a consequence, inferring cause and effect relationships from purely observational data, known as causal discovery or causal learning, has drawn much attention.

Various causal discovery methods have been proposed in the past decades, including constraint-based methods, structural equation models-based methods, and time series-based methods. Among these methods, time series-based methods, e.g., Granger causality, are relatively well-established as the temporal information excludes the case that effects happen before causes. Many of the existing time series-based methods assume that the data are measured at the right frequency; however, in practice the sampling frequency of the data is often lower than the true causal frequency. In this thesis, we consider learning high-resolution causal relationships at the causal frequency from subsampled time series. Existing methods suffer from the identifiability problems: under the Gaussianity assumption of the data, the solutions are generally not unique. We prove that, however, if the noise terms are non-Gaussian, the underlying model is identifiable from subsampled time series under mild conditions. We then propose an Expectation-Maximization approach and a variational inference approach to recover causal relations from subsampled data.

More recently, researchers began to touch upon implications of causal models for machine learning tasks such as semi-supervised learning and domain adaptation. In this thesis, we develop causally-inspired learning methods for domain adaptation in both multi-source and single-source settings. In particular, we use causal models to represent the relationship between the features and labels, and consider possible situations where different modules of the causal model change with the domain. In each situation, we investigate what knowledge is appropriate to transfer and find the optimal target-domain hypothesis. Furthermore, we propose methods to correct distribution shift in the general situation where the marginal distribution of features and conditional distribution of labels given features both change, under the assumption that labels are causes for features. We provide theoretical analysis and empirical evaluation on both synthetic and real-world data to show the effectiveness of our methods.

Contents

Contents x

List of Figures xiii

1 Introduction **2**

1.1 Outline 7

1.2 Contributions 8

2 Causal Modeling **10**

2.1 Graph Notations 10

2.2 Causal Bayesian Networks 13

2.3 Structural Equation Models 18

2.4 Time Series Models 20

3 Causal and Causally-inspired Learning **22**

3.1 Causal Learning 23

3.1.1 Constraint-based Approach 23

3.1.2 Score-based Approach 26

3.1.3 SEM-based Approach 29

3.1.4 Time Series-based Approach 34

3.2 Causally-inspired Learning 36

3.2.1 Causally-inspired Semi-supervised Learning 36

3.2.2 Causally-inspired Domain Adaptation 37

4 Discovering Temporal Causal Relations from Subsampled Data **40**

4.1 Introduction 41

4.2 Granger Causality and Its Extension with Instantaneous Effects . 44
4.3 Identifiability of the Causal Relations from Subsampled Data . . . 44
4.3.1 Effect of Subsampling (Systematic Sampling) 45
4.3.2 Identifiability of the Causal Relations at the Causal Frequency 47
4.3.3 Relation to Granger Causality with Instantaneous Effects . 51
4.4 Estimating the Temporal Causal Relations from Subsampled Data 52
4.4.1 Parameter Estimation via EM algorithm 53
4.4.2 Mean Field Approximation 58
4.4.3 Determination of the Subsampling Factor k 59
4.5 Experimental Results 59
4.5.1 Simulated Data 60
4.5.2 Real Data 62
4.6 Conclusion 65

5 Multi-source Domain Adaptation: A Causal View 66
5.1 Introduction 67
5.2 Possible Domain Adaptation Situations and Their Solutions . . . 69
5.3 Domain Adaptation with Independently Changing P_Y and $P_{X|Y}$. 73
5.3.1 The model: Target Conditional as a Linear Mixture of Source Conditionals 74
5.3.2 Parameter Estimation by Reproducing the Target Feature Distribution 78
5.3.3 Construction of the Target Classifier 80
5.3.4 Special case: Distribution Weighted Hypothesis Combination 82
5.4 Experiments 83
5.4.1 Simulations 83
5.4.2 Pseudo-real data (Sentiment Analysis) & Real-world Data (Object Recognition) 86
5.5 Related Work and Discussions 87
5.6 Practical Issues in Estimating β_{ij} in Sec. 5.3.2 88
5.7 Details of `genar_model` in Sec. 5.3.3 91
5.8 Conclusion 92

6 Domain Adaptation with Conditional Transferable Components 94

6.1 Introduction . . . 95

6.2 Conditional Transferable Components . . . 97

6.2.1 Conditional Invariant Components . . . 98

6.2.2 Location-scale Conditional Transferable Components . . . 102

6.2.3 Target Information Preservation . . . 103

6.2.4 Reparameterization . . . 104

6.2.5 Optimization . . . 105

6.3 Theoretical Analysis . . . 105

6.4 Relation to IC-type Methods . . . 108

6.5 Experiments . . . 110

6.5.1 Simulations . . . 110

6.5.2 Object Recognition . . . 111

6.5.3 Cross-domain Indoor WiFi Localization . . . 113

6.6 Proofs . . . 114

6.6.1 Proof of Theorem 6 . . . 114

6.6.2 Proof of Lemma 2 . . . 114

6.6.3 Proof of Theorem 7 . . . 115

6.6.4 Proof of Theorem 8 . . . 116

6.7 Conclusion . . . 122

7 Conclusions 123

References 125

List of Figures

2.1 A directed acyclic graph (DAG). 12
2.2 Two DAGs that are Markov equivalent to the DAG in Fig. 2.1. . 14
2.3 Three DAGs in which $P(X_2|do(X_1 = x_1)) = P(X_2|X_1 = x_1)$. (a)-(c) shows the original causal graphs, and (d)-(f) shows the corresponding interventional graphs in which X_1 is set to a specific value x_1. It can be seen that the interventional graph is identical to the original graph. 16
2.4 Three DAGs in which $P(X_2|do(X_1 = x_1)) \neq P(X_2|X_1 = x_1)$. (a)-(c) shows the original causal graphs, and (d)-(f) shows the corresponding interventional graphs. It can be seen that the interventional graph and the original graph are different. 18
2.5 A directed acyclic graph (DAG) for time series modeling. 21

3.1 A pattern returned by the SGS or PC algorithm. 26

3.2 Illustration of the identifiability conditions for the LiNGAM model. The two rows show the estimated linear function and noises in the causal and anti-causal directions when X and U_Y follow Gaussian distribution and uniform distribution, respectively. (a1) & (a2) show the scatter plot of X and Y and the estimated function in the causal direction $X \to Y$. (b1) & (b2) show the corresponding scatter plot of X and estimated residuals U_Y. (c1) & (c2) show the scatter plot of Y and X and the estimated function from Y to X. (d1) & (d2) show the corresponding plot of Y and residuals U_X. It can be seen that, when both X and U_Y are Gaussian, the residual by regression from Y to X is independent from Y. However, when both X and U_Y are uniformly distributed, the residual by regression from Y to X is dependent on Y. 31

3.3 Time series with hidden confounders. 35

4.1 Recovery of the causal-frequency data using the proposed EM method and the traditional methods: (a) The recovery results of the proposed NG-EM method ($PSNR = 13.4$); (b) The recovery results of the BFL method ($PSNR = 7.52$). 63

4.2 The contour plot of the negative log-likelihood function with repect to the two off-diagonal elements of $\hat{\mathbf{A}}$: (a) negative log-likelihood function of the Gaussian model computed on super-Gaussian data, (b) negative log-likelihood function of the non-Gaussian model computed on super-Gaussian data, (c) negative log-likelihood function of the Gaussian model computed on sub-Gaussian data, (d) negative log-likelihood function of the non-Gaussian model computed on sub-Gaussian data. 64

5.1 Possible situations of domain adaptation. W_s and V_s are domain-specific selection variables assumed to be independent, leading to changing P_{XY} across domains. (a) Covariate shift: P_X is changed by W_s, but $P_{Y|X}$ does not change. (b) W_s and V_s change P_X and $P_{Y|X}$, respectively. (c) Target shift: W_s changes P_Y, with $P_{X|Y}$ unchanged. (d) W_s and V_s change P_Y and $P_{X|Y}$, respectively. In the first two situations, we consider X as a cause for Y, whilst in the last two situations, Y is a cause of X. 70
5.2 Simulated data with three source domains and one target domain in one replication. 83
5.3 Boxplot of misclassification rate of each method on simulated data (50 replications). 84
5.4 The eigenvalues of Q_j (see (5.4)) in two random runs, one for $n = 3$, the other for $n = 5$. In both situations, one can see that $\text{rank}(Q_j) = 3$. 85
5.5 Local quadractic approximation for the penalty term r^s given in (5.18). 90

6.1 (a) Graphical representation of CIC. Here *domain* denotes the domain-specific selection variable. X^{ci} denotes the components of X whose conditional distribution, $P(X^{ci}|Y)$, is domain-invariant. We assume that X^{ci} can be recovered from X as $\mathcal{T}(X)$. $X^{\perp}$ denotes the remaining components of X; it might be dependent on Y given the domain, and when estimating X^{ci}, we would like such dependence to be as weak as possible so that X^{ci} contains as much information about Y as possible. (b) CTC, where $P(X^{ct}|Y)$ differs only in the location and scale across different domains for each value of Y. 98
6.2 Toy data to illustrate the difference between DIP and CIC: (a) The source domain data; (b) The target domain data. 108
6.3 Performance comparison on simulated data: (a) Classification error w.r.t. class ratio; (b) Classification error w.r.t. dimension d. . 109

Nomenclature

Abbreviations

EM	expectation maximization
DAG	directed acyclic graph
AR	autoregressive
BN	Bayesian network
CBN	causal Bayesian network
SEM	structural equation model
RCT	randomized controlled experiments
PC	Peter-Clark
GES	greedy equivalence search
DA	domain adaptation
SSL	semi-supervised learning
MCMC	Markov chain Monte Carlo
ICA	independent component analysis
LS	location-scale
MMD	maximum mean discrepancy
IC	invariant components
CIC	conditional invariant components
CTC	conditional transferable components
RKHS	reproducing kernel Hilbert space