

Faculty of Engineering and Information Technology
University of Technology, Sydney

**Learning from Heterogeneous Data
by Bayesian Networks**

A thesis submitted in partial fulfillment of
the requirements for the degree of
Doctor of Philosophy

by

Yin Song

April 2014

CERTIFICATE OF AUTHORSHIP/ORIGINALITY

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.

Signature of Candidate

To My Parents, Parents-in-law and Lu

Acknowledgments

First and foremost, I would like to express the deepest appreciation to my supervisor, Professor Longbing Cao, for his professional guidance, persistent help and continuous support throughout my Ph.D study and research. Most importantly, I learned from him many philosophical principles and disciplines that are not only beneficial for academic research, but also apply to many situations in life. I feel very lucky to have him as my advisor and I sincerely hope that we will remain mentor-and-student relationship for many years to come.

I would like to thank all the teachers in Advanced Analytics Institute (AAI), without their generous support this dissertation would not have been possible. I would particularly like to thank Dr. Jian Zhang for his patient guidance, scientific advice and insightful discussion.

I would like to thank all the other colleagues at AAI, for their selfless support on my research, my life and the completion of this dissertation.

My research has also benefited tremendously from various collaborations in other academic and industrial institutions. I would particularly like to thank Dr. Morgan Sangeux in the Royal Children Hospital, Melbourne, for his strong support on imparting me the domain knowledge of clinical gait analysis and assistant of collecting the data, when we collaborated to do a pilot project there. I would also like to thank to Dr Yanchang Zhao and Huaifeng Zhang in Department of Human Services, Canberra, for their great help in both work instruction and living guidance.

Finally, I would like to thank my family, my wife, my parents and parents-

ACKNOWLEDGMENTS

in-law. Without their understanding, support and encouragement, completing this dissertation would be impossible.

Yin Song

Oct. 2013 @ UTS

Contents

| | |
|---|------------|
| Certificate | i |
| Acknowledgment | iii |
| List of Figures | ix |
| List of Tables | xi |
| Abstract | xii |
| | |
| 1 Introduction | 1 |
| 1.1 Heterogeneous Data | 2 |
| 1.2 Why Learning with Bayesian Networks? | 5 |
| 1.3 Research Goals, Issues and Overview of the Thesis | 7 |
| | |
| 2 Background and Related Work | 12 |
| 2.1 Background | 12 |
| 2.1.1 Bayesian Networks | 12 |
| 2.1.2 Inference and Learning | 14 |
| 2.1.3 Hidden Markov Models: An Example of BNs | 15 |
| 2.2 Related Work | 19 |
| 2.2.1 Sequential Data Modeling | 19 |
| 2.2.2 Clinical Gait Data | 21 |
| 2.2.3 Social Recommendation Data | 22 |
| 2.2.4 Sequence Anomaly Detection | 25 |
| 2.3 Discussions | 26 |
| | |
| 3 The Latent Dirichlet Hidden Markov Model | 29 |
| 3.1 Introduction | 30 |

| | | |
|----------|---|-----------|
| 3.2 | Problem Statement | 32 |
| 3.3 | The Proposed Model | 33 |
| 3.3.1 | The Graphical Model | 33 |
| 3.3.2 | Learning the Model | 34 |
| 3.3.3 | The E step: Variational Inference of Latent Variables | 36 |
| 3.3.4 | The M Step: Estimation of Hyper-parameters | 42 |
| 3.4 | Empirical Study | 44 |
| 3.4.1 | Sequential Behavior Modeling | 44 |
| 3.4.2 | Sequence Classification | 50 |
| 3.5 | Summary | 51 |
| 4 | The Correlated Static-dynamic Model | 53 |
| 4.1 | Introduction | 54 |
| 4.2 | Problem Statement | 58 |
| 4.3 | Proposed Model | 59 |
| 4.3.1 | Motivation | 59 |
| 4.3.2 | The Correlated Static-Dynamic Model | 59 |
| 4.3.3 | The Parameters of the CSDM | 61 |
| 4.3.4 | Learning the CSDM | 62 |
| 4.4 | Empirical Study | 68 |
| 4.4.1 | Experimental Settings | 69 |
| 4.4.2 | Experimental Results | 71 |
| 4.5 | Summary | 75 |
| 5 | The Joint Interest-social Model | 77 |
| 5.1 | Introduction | 77 |
| 5.2 | Problem Statement | 80 |
| 5.2.1 | An Illustrative Example | 80 |
| 5.2.2 | Problem Formalization | 81 |
| 5.3 | The Joint Interest-social Model (JISM) | 81 |
| 5.4 | Learning and Prediction | 82 |
| 5.4.1 | Variational EM Learning | 82 |

| | | |
|----------|--|------------|
| 5.4.2 | Preference Prediction | 92 |
| 5.4.3 | Discussions | 93 |
| 5.5 | Empirical Studies | 93 |
| 5.5.1 | Data Sets | 94 |
| 5.5.2 | Evaluation Metrics | 94 |
| 5.5.3 | Comparison with State-of-the-art Methods | 94 |
| 5.5.4 | Performance Study on Varying the Properties of Users | 100 |
| 5.5.5 | Visualization of Some Interesting results | 100 |
| 5.6 | Summary | 103 |
| 6 | Enhance the Sequence Anomaly Detection | 108 |
| 6.1 | Introduction | 109 |
| 6.2 | Model-based Anomaly Detection and Its Limitations | 112 |
| 6.2.1 | The Anomaly Detection Algorithm | 112 |
| 6.2.2 | Limitations: Theoretical Analysis | 113 |
| 6.3 | How to Enhance the Discriminative Power | 114 |
| 6.3.1 | Objective Function | 115 |
| 6.3.2 | Proposed Feature Extractor | 116 |
| 6.4 | Proposed Implementation Framework | 116 |
| 6.4.1 | Phase 1: Feature Extraction | 118 |
| 6.4.2 | Phase 2: Learning of the Optimal Linear Classifier | 120 |
| 6.4.3 | Phase 3: Anomaly Detection | 122 |
| 6.5 | Experimental Settings | 123 |
| 6.5.1 | Data Sets | 123 |
| 6.5.2 | Comparative Algorithms | 126 |
| 6.5.3 | Performance Measures | 128 |
| 6.6 | Experimental Results | 128 |
| 6.6.1 | Synthetic Data | 128 |
| 6.6.2 | Real-world Data | 130 |
| 6.7 | Summary | 133 |

| | |
|---|------------|
| 7 Conclusions and Future Work | 134 |
| 7.1 Conclusions | 134 |
| 7.2 Future Work | 137 |
| Appendix | 138 |
| A Appendix for Chapter 3 | 139 |
| A.1 Distributions | 139 |
| A.1.1 Dirichlet Distribution | 139 |
| A.1.2 Multinomial Distribution | 139 |
| A.2 Variational Inference | 140 |
| A.2.1 The FF Form | 140 |
| A.2.2 The PF Form | 142 |
| B Appendix for Chapter 4 | 146 |
| B.1 The Phases of a Gait Cycle | 146 |
| B.2 The Full Decision tree | 147 |
| C Appendix for Chapter 5 | 149 |
| C.1 Distributions | 149 |
| C.1.1 Multivariate Gaussian Distribution | 149 |
| C.1.2 Bernoulli distribution | 149 |
| C.2 Proofs Related to the Lower Bound of the Log-likelihood . . . | 150 |
| C.3 Proofs Related to the E and M Steps | 154 |
| D Appendix for Chapter 6 | 156 |
| D.1 Proof of the Transformation | 156 |
| D.2 Proof of the Approximately Optimal Feature Extractor | 156 |
| D.3 Theoretical Comparison of Performance | 158 |
| E List of My Publications | 160 |
| Bibliography | 162 |

List of Figures

| | | |
|-----|--|----|
| 1.1 | A Toy Example. | 3 |
| 1.2 | Roadmap of the Research Activity | 9 |
| 2.1 | A Toy Example of Bayesian Networks | 14 |
| 2.2 | The Graphical Model of HMMs | 16 |
| 2.3 | The Graphical Model of HMMs. | 20 |
| 2.4 | The Graphical Model of LDA. | 21 |
| 2.5 | The Graphical Model of VBHMMs. | 22 |
| 2.6 | The Graphical Model of HMMV. | 23 |
| 3.1 | The Graphical Model for the LDHMMs. | 35 |
| 3.2 | The graphical models of variational distributions. | 38 |
| 3.3 | Log-likelihood Results on the Entree Data Set | 47 |
| 3.4 | Log-likelihood Results on the MSNBC Data Set | 47 |
| 3.5 | Comparison of Training Time. | 48 |
| 3.6 | The Hinton Diagrams for Parameters. | 49 |
| 4.1 | The Physical Examination Process. | 55 |
| 4.2 | The 3D Gait Analysis System | 56 |
| 4.3 | Example Gait Curves for One Patient with 6 Trials | 57 |
| 4.4 | The Graphical Model of the CSDM | 61 |
| 4.5 | Log-likelihood for the CSDM against the iteration numbers. | 72 |
| 4.6 | The Graphical Model for the Baseline Algorithm. | 72 |
| 4.7 | The Decision Tree to Predict Gait Patterns. | 75 |

| | | |
|-----|--|-----|
| 4.8 | Representative Gaits for Gait Pattern 1-4. | 76 |
| 5.1 | The Graphical Model of the JISM. | 83 |
| 5.2 | Comparison of In-matrix Performance (40%) | 97 |
| 5.3 | Comparison of In-matrix Performance (60%) | 98 |
| 5.4 | Comparison of In-matrix Performance (80%) | 99 |
| 5.5 | Performance Study against # of Ratings. | 101 |
| 5.6 | Performance Study against # of Links. | 102 |
| 5.7 | The Interest/Social Contribution. | 104 |
| 5.8 | The Clustering of Artists. | 105 |
| 6.1 | Some examples of Sequential Data. | 111 |
| 6.2 | The Flow Chart and Algorithm of the Proposed Framework . | 117 |
| 6.3 | AUC vs # of Training Sequences. | 129 |
| 6.4 | AUC vs Mean Sequence Lengths. | 130 |
| 6.5 | AUC vs HMMs | 130 |
| 6.6 | AUC vs Ratios. | 131 |
| B.1 | The Phases of a Gait Cycle. | 147 |

List of Tables

| | | |
|-----|---|-----|
| 1.1 | Research Issues in Each Chapter. | 10 |
| 3.1 | An Example of Sequential Behaviors | 32 |
| 3.2 | The Codebook of Navigation Operations | 45 |
| 3.3 | The Experimental Results of the Real Data Sets | 52 |
| 4.1 | An Excerpt Data Set from the Static Data | 56 |
| 4.2 | The Parameters for the Synthetic Data | 70 |
| 4.3 | Description of the Static Data | 70 |
| 4.4 | The Comparison of the Log-likelihoods | 73 |
| 5.1 | Notations in the JISM model | 106 |
| 5.2 | Detailed comparison of in-matrix Prediction. | 107 |
| 5.3 | Detailed comparison of out-of-matrix Prediction | 107 |
| 6.1 | Some Sample Data of Operating System Call Traces. | 110 |
| 6.2 | Parameters of the HMMs. | 124 |
| 6.3 | The Details of the Real Data Sets | 127 |
| 6.4 | The Experimental Results of the Real Data Sets | 132 |

Abstract

Non-i.i.d. data breaks the traditional assumption that all data points are independent and identically distributed. It is commonly seen in a wide range of application domains, such as transactional data, pattern recognition data, multimedia data, biomedical data and social media data. Two challenges of learning with such data are the existence of strong *coupling relationships* and *mixed structures (heterogeneity)* in the data. This thesis mainly focuses on learning from *heterogeneous* data, which refers to the non-i.i.d. data with mixed structures. To cater for the learning from such heterogeneous data, this thesis presents a number of algorithms based on Bayesian networks (BNs) that provide an effective and efficient method for representation of heterogeneous structures. A wide spectrum of non-i.i.d. data with different heterogeneity is studied. The heterogeneous data investigated in this thesis includes sequential data of unequal lengths, biomedical data mixed with time series and multivariate attributes, and social media data with both user/user friendship networks and user/item preference matrix. Specifically, for modeling a database of sequential behaviors with different lengths, latent Dirichlet hidden Markov models (LDHMMs), are designed to capture the dependent relationships in two levels (i.e., sequence-level and database-level). To learn the parameters of the model, we propose a variational EM-based algorithm. The learned model achieves substantial or comparable improvement over the-state-of-the-art models on predictive tasks, such as predicting unseen sequences and sequence classification. For learning miscellaneous data in clinical gait analysis, whose data consists of both sequential data and

multivariate data, a correlated static-dynamic model (CSDM) is constructed. An EM-based framework is applied to estimate the model parameters and some intuitive knowledge can be extracted from the model as by-products. Then, for learning more complicated social media data that records both the user/user friendship networks and user/item preference (rating) matrix in social media, we propose a joint interest-social model (JISM). We approximate the lower bound of the likelihood of the observed user/user and user/item interaction data and propose an iterative approach to learn the model parameters under the variational EM framework. The learned model is then used to predict unknown ratings and generally outperforms other comparison methods. Besides the above pure BNs-based models, we also propose a hybrid approach in the context of the sequence anomaly detection problem. This is because the estimation of the parameters of pure BNs-based model usually falls into local minimums, which may further generate inaccurate results for the sequence anomaly detection. Thus, we propose a model-based feature extractor combined with a discriminative classifier (i.e., SVM) to overcome the above issue, which is theoretically proved to have better performance in terms of Bayes error. The empirical results also support our theoretical proof. To sum up, this dissertation provides a novel perspective from Bayesian networks to harness the heterogeneity of non-i.i.d. data and offers effective and efficient solutions to learning such heterogeneous data.