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-

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

Yin Song
Oct. 2013 @ UTS

Contents

List of Figures

List of Tables

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.