

UNIVERSITY OF TECHNOLOGY SYDNEY
Faculty of Engineering and Information Technology

Mining Heterogeneous Enterprise Data

by

Xinxin Jiang

A THESIS SUBMITTED IN PARTIAL FULFILMENT
OF THE REQUIREMENTS FOR THE DEGREE

Doctor of Philosophy

Dissertation Directed by

Principal Supervisor Professor Chengqi Zhang and Co-Supervisor Doctor Guodong Long

Centre for Artificial Intelligence, FEIT

Sydney, Australia

2018

Certificate of Authorship/Originality

I, Xinxin Jiang declare that this thesis, submitted in fulfilment of the requirements for the award of Doctor of Philosophy, in the Faculty of Engineering and Information Technology at the University of Technology Sydney.

This thesis is wholly my own work unless otherwise reference or acknowledged. In addition, I certify that all information sources and literature used are indicated in the thesis. This document has not been submitted for qualifications at any other academic institution. This research is supported by the Australian Government Research Training Program.

© Copyright 2018 Xinxin Jiang

Production Note:
Signature removed
prior to publication.

Dedication

To my dear parents, parents-in-law, husband and daughter

Acknowledgements

First and foremost, I would like to express my earnest thanks to my supervisors, Professor Chengqi Zhang and Doctor Guodong Long, who have provided tremendous support and guidance for my PhD study and research. Prof Zhang provided me with an opportunity to study in the stimulating and interactive Centre for Artificial Intelligence, where I met and learned a lot from many smart and talented people. I have benefitted significantly from his generous help and invaluable suggestions on my research. I would like to thank Doctor Guodong Long for his continual guidance and supervision during my PhD study. He has always encouraged me to think about and explore my research interests. His creativeness and enthusiasm in solving challenging problems have greatly inspired my work. Without their professional guidance and consistent help, this dissertation would not have been possible.

I would also like to thank all the people that had a positive influence on my day-to-day enjoyment of research. My teachers, my leaders, my officemates, my friends, past and present: Prof Longbing Cao, Prof Ivor Tsang, Prof Guandong Xu, Mr Yongjun Lin, Mr Xiaorong Huang, Dr Shirui Pan, Dr Ling Chen, Dr Peng Zhang, Dr Jing Jiang, Dr Wei Liu, Dr Xueping Peng, Dr Can Wang, Dr Qinzhe Zhang and Dr Qin Zhang. They are the ones who have given me great support during both joyful and stressful times, and the ones to whom I will always be thankful.

Finally, I would like to thank my family, my husband Charles, my daughter Carolyn, my parents and parents-in-law for their unconditional encouragement and support. No words could possibly express my deepest gratitude for their endless love, self-sacrifice and unwavering help. To them, I dedicate this dissertation.

Xinxin Jiang
Sydney, Australia, 2018.

ABSTRACT

Mining Heterogeneous Enterprise Data

by

Xinxin Jiang

Enterprise data is complicated real-world data that is shared by the users of an organization, generally across departments and geographic regions. Heterogeneity is becoming one of the key characteristics inside enterprise data, because the current nature of globalization and competition stress the importance of leveraging huge amounts of enterprise accumulated data, according to various organizational processes, resources and standards. Effectively mining heterogeneous enterprise data is challenging. First, enormous amounts of enterprise data are being amassed from a greater variety of sources, which means traditional analysis methods are becoming less efficient and less effective. Second, enterprise data is becoming more and more heterogeneous. Primarily relate to social and technological trends and pertain to the shift in customer needs and expectations, today's enterprise data have more structures, greater complexity, and more intricate correlations than in years past. Third, learning tasks in enterprise applications have become more complex. Enterprise requirements are placing tighter constraints on data analysis, such as imbalanced data, class distributions, and multiple learning tasks, etc. Therefore, effectively deriving meaningful insights from complex large-scaled heterogeneous enterprise data poses an interesting, but critical, challenge.

The aim of this thesis is to investigate the theoretical foundations of mining heterogeneous enterprise data in light of the above challenges and to develop new algorithms and frameworks that are able to effectively and efficiently consider heterogeneity in four elements of the data: objects, events, context, and domains.

Objects describe a variety of business roles and instruments, for example, the or-

organisations, departments, people, products, or services involved in business systems. Object heterogeneity means that object information at both the data and structural level is heterogeneous. The inherent complexity of heterogeneous objects, and the dynamic nature of business processes, means that learning heterogeneous business objects in serial architectures is either computationally infeasible or ineffective because of heavy pre-processing. However, the cost-sensitive hybrid neural network (Cs-HNN) proposed in this thesis leverages parallel network architectures and an algorithm specifically designed for minority classification to generate a robust model for learning heterogeneous objects. Results from experiments with real-world data indicate that Cs-HNN demonstrates superior performance over baseline procedures.

Events trace an object's behaviours or the sequence of an object's activities. Event heterogeneity reflects the level of variety in business events and is normally expressed in the type and format of features. Mining event heterogeneity aims to build effective and efficient mining models by considering the heterogeneous event-related factors. The natural complexity of event heterogeneity in real-world business means that traditional pattern mining approaches tend to be insufficient. Most models are based on the assumption of homogeneity, but real-world events, activities, and behaviours are more often connected by heterogeneous features types in complex ways. These connections carry critical information for mining fruitful results. Practical and efficient sequential pattern mining approaches are needed to overcome these problems. The approach proposed in this thesis focuses on fleet tracking as a practical example of an application with a high degree of event heterogeneity. Complex fleet rental event patterns are discovered by combining heterogeneous features with measurement algorithms to provide valuable insights for business. The results from experiments on real-world datasets verify the effectiveness of the approach.

Context describes the environment and circumstances surrounding objects and events. Context heterogeneity reflects the degree of diversity in contextual features. Mining context heterogeneity aims to design algorithms to efficiently analyse complicated, context-aware correlations in business prospects. Again, the complexity inherent in heterogeneous contexts and the lack of a straightforward way to capture

complicated context-aware factors both present significant challenges to heterogeneous data mining. However, the coupled collaborative filtering (CCF) approach proposed in this thesis is able to provide context-aware recommendations by measuring the non-independent and identically distributed (non-IID) relationships across diverse contexts using the inter-item, intra-context, and inter-context correlations between items, users and context-aware factors as the basis for the coupled similarity calculations. The results form a set of experiments that compare and contrast a range of baseline models to demonstrate the effectiveness of this approach.

Domains are the sources of information and reflect the nature of the business or function that has generated the data - for example, different industries, or a sales function vs a supply function. As with the previous types of heterogeneity, domain heterogeneity is reflected in the number of sources the data has been derived from and manifests in the both the data type and format. Mining domain heterogeneity aims to effectively and efficiently model the correlations among real-world heterogeneous domains. However, cross-domain deep learning (Cd-DLA) provides a potential avenue to overcome the complexity and nonlinearity of heterogeneous domains. The approach presented in this thesis learns the correlations within a domain, across several domains, and between time-series data, within a parallel multi-task learning architecture. In a setting that uses multiple financial markets to represent heterogeneous domains, Cd-DLA is able to analyse complex domain-related correlations by capitalising on attention mechanisms and a series of recurrent neural networks (RNNs), and then predict market trends in the next trading window. Experimental results on 10 years of financial data prove that this approach produces more accurate predictions than other baselines.

Each of the approaches, algorithms, and frameworks for heterogeneous enterprise data mining presented in this thesis outperform the state-of-the-art methods in a range of backgrounds and scenarios, as evidenced by a theoretical analysis, an empirical study, or both. All outcomes derived from this research have been published or accepted for publication, and the follow-up work has also been recognised, which demonstrates scholarly interest in mining heterogeneous enterprise data as a

research topic. However, despite this interest, heterogeneous data mining still holds increasing attractive opportunities for further exploration and development in both academia and industry.

Dissertation directed by Professor Chengqi Zhang and Doctor Guodong Long
Faculty of Engineering and Information Technology

List of Publications

The papers of my PhD research that have been submitted, accepted, and published follow.

Conference papers

- C-1. **X. Jiang**, S. Pan, G. Long, J. Chang, J. Jiang, and C. Zhang, Cost-sensitive hybrid neural networks for heterogeneous and imbalanced data, in *International Joint Conference on Neural Networks (IJCNN)*, 2018. IEEE, 2018. (full paper accepted on 15th March 2018)
- C-2. **X. Jiang**, S. Pan, J. Jiang, and G. Long, Cross-domain deep learning approach for multiple financial market prediction, in *International Joint Conference on Neural Networks (IJCNN)*, 2018. IEEE, 2018. (full paper accepted on 15th March 2018)
- C-3. **X. Jiang**, W. Liu, L. Cao, and G. Long, Coupled collaborative filtering for context-aware recommendation, in *AAAI*, 2015, pp. 41724173.
- C-4. **X. Jiang**, X. Peng, and G. Long, Discovering sequential rental patterns by fleet tracking, in *International Conference on Data Science*. Springer, 2015, pp. 4249.

Journal papers

- J-1. **X. Jiang**, S. Pan, G. Long, F. Xiong, J. Jiang, and C. Zhang, Cost-sensitive parallel learning framework for insurance intelligence operation, *IEEE Transactions on Industrial Electronics*, 2018. (full paper accepted on 9th September 2018)

Contents

Certificate	ii
Dedication	iii
Acknowledgments	iv
Abstract	v
List of Publications	ix
Abbreviation	xiv
List of Figures	xv
List of Tables	xvii
1 Introduction	1
1.1 Background	1
1.1.1 Mining heterogeneous data	1
1.1.2 Types of heterogeneity in enterprise data	3
1.1.3 Challenges with mining heterogeneous enterprise data	8
1.2 Research objectives	15
1.2.1 Object heterogeneity	15
1.2.2 Event heterogeneity	16
1.2.3 Context heterogeneity	16
1.2.4 Domain heterogeneity	17
1.3 Thesis organisation	19

2	Literature Survey	21
2.1	Object heterogeneity	22
2.1.1	Heterogeneous business objects	22
2.1.2	Hybrid learning approach	24
2.1.3	Minority classification	26
2.2	Event heterogeneity	29
2.2.1	Event heterogeneity and utility-based pattern mining	29
2.2.2	Attention neural networks	31
2.3	Context heterogeneity	33
2.3.1	Context-aware algorithms	33
2.3.2	Non-independent and identical distribution	35
2.4	Domain heterogeneity	36
2.4.1	Cross-domain algorithms	36
2.4.2	Multi-task and transfer learning	37
3	Object Heterogeneity	41
3.1	Introduction	42
3.2	Preliminaries	45
3.3	A Cs-HNN for heterogeneous and imbalanced data	47
3.3.1	An HNN for heterogeneous data	48
3.3.2	Imbalanced cost-sensitive classification	50
3.3.3	Learning optimal parameters	52
3.3.4	Gradient computation with backpropagation	54
3.4	Experiments and evaluation	56
3.4.1	Datasets and experimental settings	56

3.4.2	Experimental results	59
4	Event Heterogeneity	63
4.1	Introduction	64
4.2	Preliminaries	66
4.2.1	Basic concepts and definitions	66
4.2.2	Technical framework	68
4.3	Rental event pattern mining	69
4.3.1	Data acquisition and item detection	69
4.3.2	Event pattern mining and behaviour analysis	70
4.4	Testing and results	71
5	Context Heterogeneity	76
5.1	Introduction	77
5.2	Preliminaries	79
5.3	CCF for context-aware recommendations	80
5.3.1	Coupled similarity calculations	81
5.3.2	Coupled similarity integrated-weight	83
5.3.3	Prediction calculation	84
5.4	Experiments	84
5.4.1	Data preparation	84
5.4.2	Metrics and comparison methods	85
5.4.3	Results	85
6	Domain Heterogeneity	88
6.1	Introduction	89

6.2 Preliminaries	92
6.2.1 Problem formalisation	92
6.2.2 Financial market analysis	94
6.2.3 Attention mechanisms	95
6.3 A cross-domain deep learning approach for multiple financial market prediction	95
6.3.1 Financial market prediction framework	96
6.3.2 Using attention in a learning approach	101
6.3.3 Learning optimal parameters	103
6.4 Experiments and evaluation	105
6.4.1 Datasets and experimental settings	105
6.4.2 Experiment results	107
7 Conclusion and Future Work	112
7.1 Conclusion	112
7.2 Future work	114
7.2.1 Business understanding and data processing	114
7.2.2 Data learning and model evaluation	115
7.2.3 Knowledge presentation	115
Bibliography	116

Abbreviation

AUC - area under curve

BiLSTM - bi-directional long short-term memory

CCF - coupled collaborative filtering

CdNN - cross-domain neural network

CNN - convolutional neural network

Cs-HNN - cost-sensitive hybrid neural network

DNN - deep neural network

DsNN - description neural network

ELM - extreme learning machines

HNN - hybrid neural network

HUSPM - high-utility sequential pattern mining

IdNN - inner-domain neural network

IID - independent and identically distributed

LSTM - long short-term memory

MAE - mean absolute error

MF - matrix factorization

MLP - multi-layer perceptron

MTL - multi-task learning

non-IID - non-independent and identically distributed

RMSE - root mean squared error

RNN - recurrent neural network

SqNN - sequence neural network

SVM - support vector machine

List of Figures

1.1	Heterogeneity by element: object, event, context, and domain	5
1.2	An example of object heterogeneity	6
1.3	An example of event heterogeneity	7
1.4	An example of context heterogeneity	8
1.5	An example of domain heterogeneity	9
1.6	Challenges associated with mining heterogeneous enterprise data . . .	11
2.1	: An example of an HNN with a two-step serial framework	25
2.2	Traditional attention mechanism	32
2.3	IID data vs non-IID data in real-world data mining	36
2.4	Knowledge transfer between heterogeneous domains	38
3.1	An example of a real-world heterogeneous dataset	43
3.2	The architecture of the Cs-HNN	47
3.3	Cs-HNN loss values with an increase in the number of epochs using the Insurance-FD dataset and an imbalance level of 5%, Cs-HNN effectively decreases the loss value.	62
4.1	Three sequential pattern mining approaches (a) general (b) location-based (c) usage-based	65
4.2	A rental pattern mining framework with fleet tracking data	68

4.3	Execution time and number of patterns for the four detection strategies	72
4.4	Comparison of the four detection strategies (a) number of patterns (b) execution time	73
4.5	Evaluation of the four detection strategies with length of patterns . .	74
5.1	Collaborative filtering (a) User ratings for items. (b) An example of context-aware recommendation. (c) Non-IID coupled relationships with contextual information.	79
5.2	Contextual coupled similarity where the context is 'location'.	80
5.3	MAE comparison for all models	86
6.1	Complex correlations among multiple financial markets	90
6.2	The attention mechanism u_t^{ij} denotes alignment score $f(x_t^{ij}, g)$	96
6.3	The architecture of the cross-domain deep learning approach in currency and stock markets prediction.	97
6.4	Cd-DLA's number of epochs: Cd-DLA effectively decreases the loss value with an increase in the number of epochs on the financial crisis dataset of the multiple financial markets from currency and stock market domains, regarding forecasts for the stock market in the United States.	111

List of Tables

1.1	The six phases of an end-to-end heterogeneous mining process	4
1.2	Research objectives for mining heterogeneous enterprise data	17
3.1	Descriptive statistics for the heterogeneous and imbalanced datasets .	46
3.2	Network settings for the Insurance-FD dataset	57
3.3	Network Settings for the MOBILE-CD Data Set	58
3.4	Evaluation with the Insurance-FD dataset	60
3.5	Evaluation with the Mobile-CD dataset	61
5.1	RMSE comparison for all models	86
6.1	Data preparation: trading indexes	106
6.2	Network settings for the financial data sets	108
6.3	Evaluation on financial crisis data set	109
6.4	Evaluation on non-crisis data set	110