

Faculty of Engineering and Information Technology
University of Technology Sydney

Big Data Analytics for Condition Based Monitoring and Maintenance

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

by

Zhibin Li

December 2020

CERTIFICATE OF ORIGINAL AUTHORSHIP

I, Zhibin Li declare that this thesis, is submitted in fulfilment of the requirements for the award of Doctor of Philosophy, in the School of Electrical and Data Engineering, Faculty of Engineering and Information Technology at the University of Technology Sydney.

This thesis is wholly my own work unless otherwise referenced 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.

Signature of Candidate

Production Note:

Signature removed prior to publication.

Acknowledgments

Foremost, I would like to express my sincere gratitude to my principal supervisor, Associate Prof. Jian Zhang, for his professional guidance, persistent help and continuous support throughout my PhD study and research.

I also deeply appreciate my co-supervisor Associate Prof. Qiang Wu, for his help and guidance in both research and living in Sydney.

Besides, I offer my regards and blessings to all of my co-workers at the lab: Yongshun Gong, Lu Zhang, Yazhou Yao, Xiaoshui Huang, Muming Zhao, Junjie Zhang, Huaxi Huang, Anan Du, Lingxiang Yao, Guofeng Mei and Litao Yu. I really enjoyed the time we spent together. I would also like to express my gratitude and appreciation to Lu Zhang for her help in proofreading this thesis, and her encouragement and support during my PhD journey.

Finally and most essentially, I would like to thank my parents. Without their encouragement, finishing this dissertation would be impossible; without them, nothing would have any value.

Zhibin Li

August 2020 @ UTS

Contents

Certificate of Original Authorship	i
Acknowledgment	ii
List of Figures	vi
List of Tables	viii
List of Publications	x
Abstract	xii
Chapter 1 Introduction	1
1.1 Background	1
1.2 Research Challenges	5
1.2.1 Sparse High-dimensional Data	5
1.2.2 Incomplete Data	6
1.2.3 Multi-source Data	6
1.2.4 Summary	7
1.3 Research Contributions	7
1.4 Thesis Structure	9
Chapter 2 Literature Review	11
2.1 Deterioration Modelling	11
2.2 Maintenance Strategy Optimisation	13
2.3 Failure Prediction	15
2.3.1 Failure Prediction with Sparse High-dimensional Data	16
2.3.2 Failure Prediction with Incomplete Data	18
2.3.3 Failure Prediction with Multi-source Data	19

2.4	Related Models	20
2.4.1	Factorisation Machines	20
2.4.2	Learning with Incomplete Data	23
2.4.3	Multi-view Learning	25
2.5	Summary	29
Chapter 3 Field-regularised Factorisation Machines for Sparse		
High-dimensional Data		30
3.1	Introduction	30
3.2	Preliminaries	32
3.3	Field-regularised Factorisation Machines	33
3.3.1	Motivation	33
3.3.2	Methods	34
3.3.3	Optimisation	36
3.4	Experiments	38
3.4.1	Data Set	39
3.4.2	Baselines and Hyper-parameter Tuning	40
3.4.3	Results and Metrics	41
3.5	Conclusion	45
Chapter 4 Missingness-pattern-adaptive Model for Incomplete		
Data		46
4.1	Introduction	47
4.2	Preliminaries	49
4.3	Missingness-pattern Adaptive Model	50
4.4	Generalisation Error Bound Analysis	53
4.5	Efficient Training Procedure	55
4.6	Proof of Convergence	56
4.7	Experiments	60
4.7.1	Linear Model	61
4.7.2	Neural Networks	63
4.8	Conclusion	70

Chapter 5	Sample-adaptive Multiple-kernel Learning for Learning with Multi-source Data	71
5.1	Introduction	71
5.2	Backgrounds	74
5.2.1	Failure Prediction of Railway Points	74
5.2.2	Preliminaries	75
5.3	Problem Description	77
5.3.1	Data Description	77
5.3.2	Problem Formulation	82
5.4	Methodology	83
5.4.1	Feature Extraction and Partition	83
5.4.2	Selecting Kernel Functions	84
5.4.3	Missingness-pattern-adaptive Multiple-kernel Learning	84
5.4.4	Sample-adaptive Multiple-kernel Learning	89
5.4.5	Optimisation	90
5.5	Experiments	93
5.5.1	Baselines, Evaluation Metrics and Parameter Setting	93
5.5.2	Results on Points-Subset Dataset	95
5.5.3	Results on Points-All Dataset	97
5.6	Conclusion	99
Chapter 6	Conclusion and Future Work	101
6.1	Conclusion	101
6.2	Future Directions of Data Analytics for CBM	102
	Bibliography	104

List of Figures

1.1	Illustration of a type of railway points.	4
1.2	An example of sparse high-dimensional data generated from a piece of maintenance log.	5
1.3	An example of multi-source data for failure prediction of railway points.	7
3.1	An example for constructing a feature vector from a sample in POINTS-3 dataset.	39
3.2	An example of labelling samples in POINTS-3 dataset.	40
3.3	Precision-recall curves with regard to POINTS-3 dataset. We drop the segments where recall is smaller than 0.1.	43
3.4	Receiver operating characteristic curves with regard to Phishing dataset.	44
4.1	When all features (x, y, z) are observable, we have an optimal separating plane in 4.1a. When only (x, y) are observable, the best separating line is the solid line in 4.1b. The projection of optimal separating plane in 4.1b is the dashed line. If we train one model for both cases, we will probably end with a compromise of them and get an inferior result.	48

4.2	The margin of a sample that only has one feature (the x dimension) is measured both in the higher-dimensional space (ρ_2) and the lower one (ρ_1). The lower-dimensional margin is larger and therefore we overestimates the margin. (Chechik, Heitz, Elidan, Abbeel & Koller 2008)	49
5.1	Workflow of our method.	78
5.2	A piece of IFMS data.	79
5.3	A piece of equipment details.	80
5.4	A piece of maintenance log.	80
5.5	A piece of movement log.	81
5.6	A piece of weather data.	81
5.7	To forecast failures in week $i+1$, we use data from week i and maintenance logs in a 35-day interval before week $i+1$	83

List of Tables

3.1	A sample of maintenance records with failures to be predicted.	34
3.2	Statistics of the datasets.	40
3.3	Comparison of LINEAR-LR, FM, FFM, FrFM-EUC and FrFM-COS. The best results are bold and the second-best are underlined	42
4.1	Summary of datasets	61
4.2	Classification accuracy (mean±std) with additional 30% entries removed for all datasets. The best results are bold and the second best are underlined.	64
4.3	Classification accuracy (mean±std) on original datasets. The best results are bold.	65
4.4	Classification accuracy (mean±std) on Sensorless Drive Diagnosis dataset. The best results are bold and the second best are underlined.	67
4.5	Classification accuracy (mean±std) on MNIST dataset. The best results are bold and the second best are underlined.	68
4.6	Classification accuracy (mean±std) on Avila dataset. The best results are bold and the second best are underlined.	69
5.1	Missing rates and dimensions of our data channels. 44% of samples are missing at least one channel.	85
5.2	Dataset summary.	93

5.3 Experiment results on Points-Subset dataset. Best results are bold and the second best are underlined. The results are reported with means and standard deviations (mean \pm std) for non-convex methods. 96

5.4 Experiment results on Points-All dataset. Best results are bold and the second best are underlined. The results are reported with means and standard deviations (mean \pm std) for non-convex methods. 98

List of Publications

Papers Published

- **Zhibin Li**, Jian Zhang, Yongshun Gong, Yazhou Yao, and Qiang Wu. *Field-wise Learning for Multi-field Categorical Data*, Advances in Neural Information Processing Systems, 2020.
- **Zhibin Li**, Jian Zhang, Qiang Wu, Yongshun Gong, Jinfeng Yi, and Christina Kirsch. *Sample adaptive multiple kernel learning for failure prediction of railway points*, In Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, pp. 2848–2856, 2019.
- **Zhibin Li**, Jian Zhang, Qiang Wu and Christina Kirsch. *Field-regularized Factorization Machines for Mining the Maintenance Logs of Equipment*, The 31st Australasian Joint Conference on Artificial Intelligence, 2018.
- Yongshun Gong, **Zhibin Li**, Jian Zhang, Wei Liu and Yu Zheng. *Online Spatio-temporal Crowd Flow Distribution Prediction for Complex Metro System*, IEEE Transactions on Knowledge and Data Engineering, 2020.
- Yongshun Gong, **Zhibin Li**, Jian Zhang, Wei Liu, and Jinfeng Yi. *Potential passenger flow prediction: A novel study for urban transportation development*. In Proceedings of the 34th AAAI Conference on Artificial Intelligence, pages 4020–4027, 2020.

- Yongshun Gong, **Zhibin Li**, Jian Zhang, Wei Liu, Yu Zheng, and Christina Kirsch. *Network-wide crowd flow prediction of Sydney trains via customized online non-negative matrix factorization*. In Proceedings of the 27th ACM International Conference on Information and Knowledge Management, pages 1243–1252, 2018.
- Yongshun Gong, **Zhibin Li**, Jian Zhang, Wei Liu, Bei Chen and Xiangjun Dong. *A Spatial Missing Value Imputation Method for Multi-view Urban Statistical Data*. In Proceedings of the 29th International Joint Conference on Artificial Intelligence, pages 1310-1316, 2020.
- Lu Zhang, Jian Zhang, **Zhibin Li**, and Jingsong Xu. *Towards Better Graph Representation: Two-branch Collaborative Graph Neural Networks for Multimodal Marketing Intention Detection*. IEEE International Conference on Multimedia & Expo. 2020.

Abstract

Condition-based Maintenance (CBM) will significantly achieve the cost-saving while monitoring the related infrastructure through the most accurate maintenance scheduling. It also increases the reliability of monitored equipment. For example, in the field of rail transport, it helps ensure trains run on time and plays a critical role in the safety of railway operation. A key prerequisite for CBM is accurate fault prediction, which can be achieved through predictive machine learning models. Although artificial intelligence and machine learning have become successes in many applications, their potentials in CBM have not been fully recognised. The growing scale and modality of railway data bring opportunities as well as challenges to machine learning models. In this thesis, three key challenges were abstracted with regard to data analytics using machine learning technics for fault prediction, resulting from the sparse high-dimensional data, the incomplete data, and the multi-source data. Then the three challenges were studied from an algorithmic point of view.

The sparse high-dimensional data commonly exist in maintenance logs, in a format of categorical variables. Normally, a sophisticated feature engineering process is required to extract the complex feature-interactions, while the high dimensionality, sparseness, and the lack of reliable domain knowledge make this process quite ad-hoc and subject to strong personal opinion/experience of each individual engineer. This thesis proposed field-regularised factorisation machines to learn the complex feature-interactions automatically from such data and evaluated the proposed method with main-

tenance logs of railway points in a railway network. Another challenge comes with the fact that real-world data are usually incomplete due to various reasons, e.g., faults in the database, operational errors or transmission faults. To address these issues, this thesis proposed a missingness-pattern-adaptive model, which adaptively adjusts the predictive function for incomplete data. Some theoretical evidence was provided to support the correctness of our model. This model was tested with several public datasets with internal missing values. Generally, the predictive task for CBM can involve data from multiple sources, such as weather conditions, sensors, and maintenance logs. For the multi-source data, this thesis proposed a sample-adaptive multiple-kernel learning algorithm to facilitate the fusion of data for the predictive task. To verify the effectiveness of this method, experiments were conducted on real-life data generated by a large-scale railway network.