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

List of Figures

List of Tables

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 maintenance 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.