Non-IID Outlier Detection with Coupled Outlier Factors

Guansong Pang

Supervisors:
Longbing Cao, Principal
Ling Chen
Huan Liu, Arizona State University

This thesis is presented as part of the requirements for the conferral of the degree:

Doctor of Philosophy

University of Technology Sydney
Faculty of Engineering and Information Technology

April 2019

To my beloved parents, Yuxiang Pang and Shangxing Li.
To my wife Lisha and my son Louis.

Certificate of Original Authorship

I, Guansong Pang, declare that this thesis, submitted in fulfillment of the requirements for the award of the degree: 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.*

Guansong Pang
April 1, 2019

Acknowledgments

I am greatly indebted to my principal supervisor, Professor Longbing Cao, one of the best mentors I have ever met. Over the past three years, Longbing spent tremendous time on teaching me how to think critically, do solid research and write 'beautiful' technical papers. He has been very nice, thoughtful and supportive to me in that he gave me large freedom to exclusively explore interesting research problems and publish high-quality papers at the early stage of my candidature while prepared me for an independent academic career in my senior candidature by involving me in different professional activities (e.g., conference/workshop organization, program committee of leading conferences), conference tutorials, student mentoring, research proposals, and industry project meetings. Longbing has been very fruitful, critical and constructive in making comments on research. I will not forget his massive advice on my designs and manuscripts, which are extremely important and helpful for sharpening my research designs and paper writings. In addition to research, Longbing has been a great life mentor of me for sharing many wonderful and undesirable experiences to encourage me when I was down and to remind me pitfalls when I was in triumph. I cannot thank him enough for all his supports.

I would like to thank my co-supervisor Dr. Ling Chen and external supervisor Professor Huan Liu for their helpful and insightful comments on both of my research designs and papers. I would like to particularly thank Ling for her nice and warm greetings on many festivals and public holidays. Ling has been always very supportive and understanding for many aspects in my research and daily life I seek suggestions and help from her. I am very grateful to Huan for not only having nice and thorough discussions over my research whenever I met him, but also suggesting interesting directions for my work, offering me great advice on job applications and career planning.

Besides my supervisors, I would also like to thank the panel members of my candidature assessment, Dr. Wei Liu, Professor Jinyan Li, Dr. Haiyan Lu, and Professor Bogdan Gabrys, for their constructive comments and great encouragement.

I thank all my friends and colleagues in the Data Science Lab at AAi for being critical and insightful to my research in our weekly research meetings and being very nice to me whenever I met them. I particularly thank: Chengzhang Zhu for many wonderful discussions about our research and careers, Songlei Jian for the nice research discussions and collaborations, Shoujin Wang and Thac Do for their companion and story sharing in our quiet lab. I am very grateful to visiting professors Defu Lian, Wenpeng Lu, and Lizhen Wang for their constructive comments to my research and great help in my career planning. I really appreciate the collaboration with Hongzuo Xu. It was a pleasure to

work with him and I learned many helpful mentoring skills from this experience. I would also like to thank the other friends in the lab who enrich my research life and/or give me consistent encouragement: Wei Wang, Liang Hu, Qi Zhang, Ke Liu, Longxiang Shi, Wenfeng Hou, Qing Liu, Usman Naseem, Yan Xing, Jingyu Shao, Frank Xu, Dr. Allen Lin, Professor Quangui Zhang, and Dr. Lei Gu. It was always a great time to celebrate every of our success in the lab and to gather together to have fun and enjoy the holidays in festivals.

I would like to express my special thanks to Professor Shengyi Jiang who introduced me to the academic community and was always there to give me advice whenever I seek help from him. I am also very thankful to Dr. Huidong Jin, Professor Kai Ming Ting, and Dr. David Albrecht for their enormous time and effort on my previous research which lays an important foundation to my PhD research. I also thank everyone who has inspired me and helped me in my research.

I would also like to thank the UTS Advanced Analytics Institute, School of Software, and Graduate Research School for their quality administration service and financial support for my domestic and international conference travels. Additionally, I thank Michele Mooney for her excellent proofreading of this thesis.

Last but not least, I wholeheartedly thank my beloved mother and father for their everlasting love, amazing encouragement and strong support throughout my life. I am also specially grateful to my brothers and sisters who are very supportive in taking care of all the family matters during my graduate study. Special thanks to my dear wife Lisha Zhang for being patient and extremely supportive during this tough period and for planning my career and making decisions on our future with me. I am so fortunate to have her in my life. I would also like to thank my son Louis (Zhizhong) Pang who colors my life with his cheerfulness, lovely and gorgeous smiles, and chaos of trouble. This thesis is dedicated to them.

Publications During Candidature

Papers that have been published or accepted for publication:

1. **Guansong Pang**, Longbing Cao, Ling Chen, and Huan Liu. "Learning Representations of Ultrahigh-dimensional Data for Random Distance-based Outlier Detection", In: *24th SIGKDD Conference on Knowledge Discovery and Data Mining (KDD-18)*. London, UK. Accepted (long presentation).

2. Songlei Jian, **Guansong Pang**, Longbing Cao, Kai Lu, and Hang Gao. "CURE: Flexible Categorical Data Representation by Hierarchical Coupling Learning". *IEEE Transactions on Knowledge and Data Engineering (TKDE)*. Accepted.

3. **Guansong Pang**, Longbing Cao, Ling Chen, Defu Lian and Huan Liu. "Sparse Modeling-based Sequential Ensemble Learning for Effective Outlier Detection in High-dimensional Numeric Data", In: *32nd AAAI Conference on Artificial Intelligence (AAAI-18)*. AAAI Press, pp. 3892-3899. New Orleans, US.

4. **Guansong Pang**, Longbing Cao, Ling Chen and Huan Liu. "Learning Homophily Couplings from Non-IID Data for Joint Feature Selection and Noise-Resilient Outlier Detection". In: *26th International Joint Conference on Artificial Intelligence (IJCAI-17)*. AAAI Press, pp. 2585-2591.

5. Songlei Jian, Longbing Cao, **Guansong Pang**, Kai Lu and Hang Gao. "Embedding-based Representation of Categorical Data by Hierarchical Value Coupling Learning". In: *26th International Joint Conference on Artificial Intelligence (IJCAI-17)*. AAAI Press, pp. 1937-1943.

6. **Guansong Pang**, Hongzuo Xu, Longbing Cao and Wentao Zhao. "Selective Value Coupling Learning for Detecting Outliers in High-Dimensional Categorical Data". In: *26th ACM International Conference on Information and Knowledge Management (CIKM-17)*, Long paper track. ACM, pp. 807-816. Singapore.

7. **Guansong Pang**, Kai Ming Ting, David Albrecht and Huidong Jin. "ZERO++: Harnessing the Power of Zero Appearances to Detect Anomalies in Large-Scale Data Sets". *Journal of Artificial Intelligence Research (JAIR)* 57, pp. 593–620, 2016.

8. **Guansong Pang**, Longbing Cao, and Ling Chen. "Outlier Detection in Complex Categorical Data by Modeling Feature Value Couplings". In: *25th International*

Joint Conference on Artificial Intelligence (IJCAI-16). AAAI Press, pp. 1902–1908, 2016. New York City, US.

9. **Guansong Pang**, Longbing Cao, Ling Chen and Huan Liu. "Unsupervised Feature Selection for Outlier Detection by Modeling Hierarchical Value-Feature Couplings." In: *2016 IEEE International Conference on Data Mining (ICDM-16)*, Long paper track. IEEE, pp. 410-419. Barcelona, Spain.

10. **Guansong Pang**, Kai Ming Ting, and David Albrecht. "LeSiNN: Detecting anomalies by identifying Least Similar Nearest Neighbours". In: *2015 IEEE 15th International Conference on Data Mining Workshops (ICDMW-15)*. IEEE, pp. 623–630, 2015.

Papers that are under review:

1. **Guansong Pang**, Longbing Cao, and Ling Chen. "Outlier Detection with Non-IID Outlier Factors". *Data Mining and Knowledge Discovery (DMKD)*. Under review.

2. **Guansong Pang** and Longbing Cao. "Building Optimal Heterogeneous Univariate Outlier Ensembles". *IEEE Transactions on Neural Networks and Learning System (TNNLS)*. Under review.

Abstract

Outliers are data objects which are rare or inconsistent compared to the majority of objects. Outlier detection is one of the most important tasks in data mining due to its wide applications in various domains, such as finance, information security, healthcare and earth science. Most existing outlier detection methods assume that the outlier factor (i.e., outlierness scoring measure) of the entities (e.g., feature values, features, data objects) in a data set is Independent and Identically Distributed (IID), but this assumption is violated by many real-world applications where the outlierness of an entity is coupled with that of some other entities, leading to the failure of detecting sophisticated outliers. This issue is intensified in more challenging environments, e.g., noisy and/or high-dimensional data sets. To address this challenge, this thesis considers three key questions: what are the coupling relations between different outlier factors? how can we effectively and efficiently model these couplings? and how can we leverage these couplings to address challenging outlier detection problems?

Our explorations result in the following four key contributions. (i) This thesis introduces a *new outlier detection task*, non-IID outlier detection in multidimensional data, which opens a new research direction for tackling real-world complex outlier detection problems. (ii) We introduce the *first architecture* for the non-IID outlier detection task, which provides principled approaches to learn the outlierness interdependence at different levels from feature values, features, to data objects. The architecture breaks down the general coupling learning into a series of important finer-grained components: basic coupling relation, coupling capacity, coupling utility, and coupling passage manners, providing feasible ways to learn sophisticated couplings between outlier factors with efficient models. (iii) We propose principled frameworks and their instantiations under the non-IID outlier detection architecture to learn different types of couplings. Supported by extensive theoretical analysis and empirical experiments on diverse real-world data sets, these designs are shown to be scalable and effective in *addressing some notoriously challenging problems*, including outlier detection in non-IID data, data with many noisy features, or high-dimensional data. (iv) This thesis also introduces a set of *seminal work* on unsupervised feature selection for outlier detection in both categorical data and numeric data, including innovative feature selection methods that capture pairwise or full feature interactions and joint feature selection and outlier detection methods. Our proposed approaches are able to effectively compute the outlierness of features, which enables outlying feature selection and substantially improves the efficacy of subsequent outlier detection on data with high dimensionality or many noisy features.

Our extensive empirical results show that the average accuracy improvement of our non-IID outlier detectors over state-of-the-art IID outlier detectors ranges from 4% up to 18% on a large collection of real-world data sets; the maximum accuracy improvement on single data sets can be more than 50%, in which stat-of-the-art IID detectors only obtain an accuracy of being nearly equivalent to a random guess. This significant accuracy improvement can have great business value, e.g., the prevention of millions of dollars loss in credit card fraud detection, enabling safer digital environments by militating malicious programs or network intrusions, or saving life by having early detection of fatal diseases. This thesis also offers much more interpretable outlier detection solutions by enabling outlier detection in highly relevant and substantially smaller feature subsets.

Contents

List of Tables

List of Figures