C02029: Doctor of Philosophy
CRICOS Code: 009469A
Subject Code: 32903
October 2019

Non-IID Representation Learning on Complex Categorical Data

Chengzhang Zhu

School of Computer Science
Faculty of Engineering and Information Technology
University of Technology Sydney
NSW - 2007, Australia

Non-IID Representation Learning on Complex Categorical Data

A thesis submitted in partial fulfilment of the requirements for the degree of

Doctor of Philosophy
in
Analytics

by

Chengzhang Zhu

to

School of Computer Science
Faculty of Engineering and Information Technology
University of Technology Sydney
NSW - 2007, Australia

October 2019

Abstract

Learning complex categorical data requires proper vector or metric representations of the intricate characteristics of that data. Existing methods for categorical data representation usually assume data is independent and identically distributed (IID). However, real-world data is often hierarchically associated with diverse couplings and heterogeneities (i.e., non-IIDness, e.g., various couplings such as value co-occurrences and attribute correlation and dependency, as well as heterogeneities such as heterogeneous distributions or complementary and inconsistent relations). Existing methods either capture only some of these couplings and heterogeneities or simply assume IID data in building their representations.

This thesis aims to deeply understand and effectively represent non-IIDness in categorical data. Specifically, it focuses on (1) modeling heterogeneous couplings within and between attributes in categorical data; (2) disentangling attribute couplings with a mixture of heterogeneous distributions; (3) hierarchically learning heterogeneous couplings; (4) integrating complementary and inconsistent heterogeneous couplings; and (5) adaptively identifying and learning dynamic couplings and heterogeneities.

Accordingly, this thesis proposes (1) a non-IID similarity metrics learning framework to model complex interactions within and between attributes in non-IID categorical data; (2) a decoupled non-IID learning framework to capture and embed heterogeneous distributions in non-IID categorical data with bounded information loss; (3) a heterogeneous metric learning method with hierarchical couplings to learn and integrate the heterogeneous dependencies and distributions in non-IID categorical data into a representation of a similarity metric; (4) an unsupervised heterogeneous coupling learning approach to integrate the complementary and inconsistent heterogeneous couplings in non-IID categorical data; and (5) an unsupervised hierarchical and heterogeneous coupling learning method to learn hierarchical and heterogeneous couplings on dynamic non-IID categorical data.

Theoretical analyses support the effectiveness of the proposed methods and bound the information loss in their generated high-quality representations. Extensive experiments demonstrate that the proposed non-IID representation methods for complex categorical data perform significantly better than state-of-the-art methods in terms of multiple downstream learning tasks and representation-quality evaluation metrics.

Author's declaration

I, *Chengzhang Zhu* declare that this thesis, is submitted in fulfilment of the requirements for the award of Doctor of Philosophy, in the *School of Computer Science*, *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.

Production Note:

SIGNATURE: Signature removed prior to publication.

Chengzhang Zhu

DATE: 22nd October, 2019

PLACE: Sydney, Australia

DEDICATION

To my son Jiawei Zhu...

Acknowledgments

First and foremost I want to thank my supervisor Prof. Longbing Cao. It has been an honor to be his Ph.D. student. He has taught me the skills required by an independent researcher, especially critical thinking and academic writing. I appreciate all his contributions of time, ideas, discussions, and paper polishing to make my Ph.D. experience impressive and stimulating. I am deeply influenced by his professionalism, hard work, and profound ideas. I am also thankful for the excellent opportunity to serve in the research community and the industry internship he has provided.

I would also like to thank Prof. Jianping Yin for his insightful comments, encouragement, and support. His optimism and enthusiasm were contagious and motivational for me, even during the difficult times in my Ph.D. pursuit.

My sincere thanks also go to all of my colleagues in the Advanced Analytics Institute, for their enthusiastic help: Dr. Guansong Pang, Dr. Trong Dinh Thac Do, Dr. Shoujin Wang, Dr. Zhenjiang Lin, Dr. Liang Hu, Qi Zhang, Qing Liu, Zhilin Zhao, Wei Wang, and Jia Xu. I appreciate our intensive discussions, in-depth collaboration, and all the fun we have had in the last four years.

I thank the visiting scholars and students in our laboratory. They broadened the horizon of my research and shared unforgettable times with me at UTS. In particular, I am grateful to Prof. Wentao Zhao, Prof. Lizheng Wang, A/Prof. Quangui Zhang, A/Prof Defu Lian, A/Prof. Lingyun Xiang, Dr. Wenpeng Lu, Dr. Xinwang Liu, Songlei Jian, Shangdong Yang, and Teng Li for their precious advice and encouragement in my Ph.D. research.

I gratefully acknowledge the funding from China Scholarship Council that made my Ph.D. work possible.

Lastly, I would like to thank my family for all their love and encouragement. Without their support, it would not be possible for me to conduct this research. In particular, I am grateful to my wife Qian Li whose spiritual and academic support are so appreciated. Thank you.

Chengzhang Zhu
University of Technology Sydney
October 2019

List of Publications

Related to the Thesis :

Published Journal Papers :

1. **Zhu, C.**, Cao, L., Liu, Q., Yin, J. and Kumar, V., 2018. Heterogeneous metric learning of categorical data with hierarchical couplings. *IEEE Transactions on Knowledge and Data Engineering*, 30(7), pp.1254-1267.
2. Zhao, W., Li, Q., **Zhu, C.**, Song, J., Liu, X. and Yin, J., 2018. Model-aware categorical data embedding: a data-driven approach. *Soft Computing*, 22, pp.3603-3619.

Published Conference Papers :

3. Zhang, Q., Cao, L., **Zhu, C.**, Li, Z. and Sun, J., 2018, July. CoupledCF: Learning explicit and implicit user-item couplings in recommendation for deep collaborative filtering. *In International Joint Conference on Artificial Intelligence* (pp. 3662-3668).
4. Zhao, X., **Zhu, C.** and Cheng, L., 2017, October. Coupled Bayesian matrix factorization in recommender systems. *In 2017 IEEE International Conference on Data Science and Advanced Analytics* (pp. 522-528). IEEE.
5. Song, J., **Zhu, C.**, Zhao, W., Liu, W. and Liu, Q., 2017, September. Model-aware representation learning for categorical data with hierarchical couplings. *In International Conference on Artificial Neural Networks* (pp. 242-249). Springer, Cham.

Submitted Manuscripts :

6. **Zhu, C.**, Cao, L., and Yin, J., 2019. Unsupervised categorical representation with heterogeneous and inconsistent couplings. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

7. **Zhu, C.**, Cao, L., Liu, Q., and Yin, J., 2019. Decoupled non-IID nominal data representation. *Journal of Machine Learning Research*.
8. **Zhu, C.**, Cao, L., and Yin, J., 2018. Similarity metrics for non-IID categorical data. *Artificial Intelligence*.
9. Li, Q., Wu, Q., **Zhu, C.**, Cao, L., and Zhang, J., 2019. Deep non-IID paragraph representation for sentiment analysis. *Pattern Recognition*. (Major Revision)

OTHERS :

Published Journal Papers :

10. Yuan, L., Liu, Q., Lu, M., Zhou, S., **Zhu, C**. and Yin, J., 2017. A highly efficient human activity classification method using mobile data from wearable sensors. *International Journal of Sensor Networks*, 25(2), pp.86-92.
11. Zhao, L., Li, K., Wang, M., Yin, J., Zhu, E., Wu, C., Wang, S. and **Zhu, C.**, 2016. Automatic cytoplasm and nuclei segmentation for color cervical smear image using an efficient gap-search MRF. *Computers in biology and medicine*, 71, pp.46-56.
12. Zhou, S., Liu, X., Liu, Q., Wang, S., **Zhu, C.** and Yin, J., 2016. Random Fourier extreme learning machine with $\ell 2$, 1-norm regularization. *Neurocomputing*, 174, pp.143-153.
13. Liu, Q., Zhou, S., **Zhu, C.**, Liu, X. and Yin, J., 2016. MI-ELM: Highly efficient multi-instance learning based on hierarchical extreme learning machine. *Neurocomputing*, 173, pp.1044-1053.

Published Conference Papers :

14. Li, Q., Wu, Q., **Zhu, C.**, and Zhang, J., 2019, September. Bi-level masked multi-scale CNN-RNN networks for short text representation *In International Conference on Document Analysis and Recognition*. IEEE.
15. Li, Q., Wu, Q., **Zhu, C.**, Zhang, J. and Zhao, W., 2019, July. An inferable representation learning for fraud review detection with cold-start problem. *In 2019 International Joint Conference on Neural Networks*. IEEE.
16. Li, Q., Wu, Q., **Zhu, C.**, Zhang, J. and Zhao, W., 2019, April. Unsupervised user behavior representation for fraud review detection with cold-start problem. *In*

Pacific-Asia Conference on Knowledge Discovery and Data Mining (pp. 222-236). Springer, Cham.

17. Wang, S., Zeng, Y., Liu, Q., **Zhu, C.**, Zhu, E. and Yin, J., 2018, October. Detecting abnormality without knowing normality: A two-stage approach for unsupervised video abnormal event detection. *In 2018 ACM Multimedia Conference on Multimedia Conference* (pp. 636-644). ACM.

18. Zhao, G., Xiang, L., **Zhu, C.** and Li, F., 2018, July. Two-stage unsupervised multiple kernel extreme learning machine. *In 2018 International Joint Conference on Neural Networks*. IEEE.

Table of Contents

List of Figures

List of Tables

Table **Page**