Concept Drift Adaptation for Real-time Prediction

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

Doctor of Philosophy
in
Computer Science

by

Yiliao Song

to

School of Computer Science, Faculty of Engineering and Information Technology
Australian Artificial Intelligence Institute

University of Technology Sydney
NSW - 2007, Australia

August 2020

CERTIFICATE OF ORIGINAL AUTHORSHIP

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

Production Note:
Signature removed prior to publication.

SIGNATURE: ______________________

DATE: August, 2020

Abstract

Concept drift refers to the phenomenon of distribution changes in a data stream. Using concept drift adaptation techniques to predict the target variable(s) of real-time data streams has gained the ever-increasing attention of researchers in recent years.

This research aims to develop a set of concept drift adaptation methods for predicting the target variable of real-time data streams. The literature review reveals two issues in the area of concept drift: i) how the concept drift problem limits the learning capability; ii) how to make adaptation in more realistic scenarios that data streams have uncertainties other than concept drift.

To address the issue i), this research discovers three root causes of limited learning capability when concept drift occurs. It is found that when concept drift occurs in a data stream, the prediction accuracy is decreased because 1) the training set contains more than one patterns so that the predictor cannot be well-learned; 2) a newly arrived data instance may present old patterns but an old instance presents the new pattern; and 3) few data instances are available when a new concept is identified at its early stage. Three concept drift adaptation methods are designed to address the three situations separately. Situation 1) is solved by developing a *fuzzy clustering-based adaptive regression* (FUZZ-CARE) approach. FUZZ-CARE can learn how many patterns exist in the training set and the membership degree of each instance belonging to each pattern; To learn the predictor with the most relevant data rather than the newest arrived data, a *segment-based drift adaptation* (SEGA) method to sequentially pick out the best segments in the training data to update the predictors. This addresses the situation 2). An *adaptive fuzzy network* (AFN) is designed to address the situation 3) through generating samples of the new concept with the previous data instances.

To address the issue ii), this research discusses the concept drift phenomenon under two scenarios that are more realistic. One is to solve the concept drift problem when data is noisy. A *noise-tolerant drift adaptation* (NoA) method is designed for handling concept drift when the data stream contains signal

noise; the other is to solve the concept drift problem when data also contains temporal dependency. A theoretical study is conducted for the regression of data streams with concept drift and temporal dependency, and based on this study, a *drift-adapted regression* (DAR) framework is established.

To conclude, this thesis not only provides a set of effective drift adaptation methods for real-time prediction, but also contributes to the development of concept drift area.

Dedication

To my loving husband and parents.

Acknowledgments

It is a memorial and exciting journey at University of Technology Sydney (UTS) for pursuing my Ph.D. degree in the past four years. I am sincerely grateful to the people who inspired and helped me in many ways.

I would like to express my foremost and deepest gratitude to my principal supervisor, Distinguished Professor Jie Lu. Without her patience and encouragement, I would not have been able to complete this Ph.D. program. She led me into a new academic research direction, and convincingly guided and encouraged me to think and work as a professional scientist. She placed considerable trust in my research ability and unconditionally support me in pursuing my own research interests. Her wisdom and immense knowledge always enlightened me to go further and deeper in my research. Her decisiveness and sharp insights continuously motivated me when I got lost or afraid about the future. Her confidence and enthusiasm inspired me to do the right thing even when the road got tough. I felt extremely honored to be guided by such a rigorous researcher as well as an enthusiastic mentor. What she taught me and what I learned from her in the past four years has benefited my Ph.D. study and will be a great treasure throughout my life.

Meanwhile, I am greatly indebted to my co-advisor, A./Professor Guangquan Zhang. He taught me step by step how to become a qualified researcher from its beginning. He always led me to the right research direction with his expert knowledge of theory and abundant research experience. Without his critical comments, I would waste my time on trivial research ideas. Discussion with him greatly improves the scientific aspect and quality of my research. He helped me to build my confidence in my research outcomes and to be hopeful when faced with any difficulty, from academic to living.

I would like to truly thank my co-advisor, A./Professor Haiyan Lu. Without her endless trust and help, I would never have this valuable opportunity to pursue my Ph.D. research in UTS. She encouraged me to participate in different research activities which have significantly broaden my research horizons. She gave me constant care over my conditions. Her kindness helped me to go through the tough period in my Ph.D. program. It is always pleasant to share all kinds of opinions with her.

During my Ph.D. period, I am very fortunate to join the Department of Computer Science at Southern University of Science and Technology as a visiting scholar, working with Prof. Xin Yao and Dr. Changwu Huang for three months. I would like to express my thankfulness to Prof. Xin Yao and Dr. Changwu Huang that they led me to the evolutionary computation area. Discussion and cooperation with them helped me to fill my knowledge blank of evolutionary computation efficiently and enlightened me on my own research at the same time.

I would like to express my thankfulness to every member of the Decision Systems & e-Service Intelligence Lab (DeSI) in the Centre for Artificial Intelligence (CAI) for their careful participation in my presentation and valuable comments for my research. It was a wonderful experience to spend four years with these dedicated researchers. I especially thank Dr. Ning Lu, Dr. Anjin Liu, Dr. Fan Dong and Dr. Feng Gu who helped me greatly to deeply understand my research problem during my Ph.D. candidature; Dan Shang, Hang Yu, Bin Wang and Bin Zhang who have shared their opinions and comments with me, Dr. Hua Zuo, Dr. Qian Zhang, Dr. Shan Xue who shared my joys and sadness, Dr. Dianshuang Wu who helped me in the living.

I genuinely thank Sue Felix and Robyn Barden for polishing the language of my publications. They are always patient to all my emails of questioning revised sentences. I have learned much about academic writing from them.

Meanwhile, I must thank Dr. Xin Wang and Dr. Bo Han for the valuable suggestions at the beginning of my Ph.D. study. Although they are working in different research areas, their persistence and passion on research impressed and inspired me. I thank all my wonderful friends, classmates and colleagues for every enjoyable moment.

Last, I would like to express my heartfelt appreciation and gratitude to my husband, parents, and families for their love and support.

List of Publications

1. Y. Song, G. Zhang, H. Lu, J. Lu, “Fuzzy Clustering-based Adaptive Regression for Drifting Data Streams", *IEEE Transactions on Fuzzy Systems*, (2019). [CORE A*]

2. Y. Song, G. Zhang, H. Lu and J. Lu, “A Fuzzy Drift Correlation Matrix for Multiple Data Stream Regression", *IEEE International Conference on Fuzzy Systems*, Glasgow, Scotland (2020). [CORE A]

3. Y. Song, G. Zhang, J. Lu, H. Lu, “A Noise-tolerant Fuzzy c-Means based Drift Adaptation Method for Data Stream Regression", *IEEE International Conference on Fuzzy Systems*, New Orleans, USA (2019). [CORE A]

4. Y. Song, G. Zhang, H. Lu, J. Lu, “A Self-adaptive Fuzzy Network for Prediction in Non-stationary Environments", *IEEE International Conference on Fuzzy Systems*, Rio de Janeiro, Brazil (2018). [CORE A]

5. Y. Song, G. Zhang, J. Lu, H. Lu, “A fuzzy kernel c-means clustering model for handling concept drift in regression", *IEEE International Conference on Fuzzy Systems*, Naples, Italy (2017). [CORE A]

6. A. Liu, Y. Song, G. Zhang, and J. Lu, “Regional concept drift detection and density synchronized drift adaptation", *in the 26th International Joint*

Conference on Artificial Intelligence , Melbourne, Australia (2017). [CORE A*]

7. J. Lu, A. Liu, Y. Song, G. Zhang, "Data-driven Decision Support under Concept Driftin Streamed Big Data", *Complex & Intelligent Systems* , (2019).

8. Y. Song, J. Lu, H. Lu, G. Zhang, "A Segment-based Drift Adaptation Method for Data Streams", *IEEE Transactions on Neural Networks and Learning Systems* , (submitted). [CORE A*]

9. Y. Song, J. Lu, H. Lu, G. Zhang, "A drift-adapted regression framework for data streams with temporal dependency", *Artificial Intelligence* , (submitted). [CORE A*]

10. F. Dong, J. Lu, Y. Song, F. Liu, G. Zhang, "A Drift Region-Based Data Sample Filtering Method", *IEEE Transactions on Cybernetics* , (submitted). [CORE A]

Table of Contents

List of Figures

Figure **Page**

List of Tables