CONCEPT DRIFT ADAPTATION FOR LEARNING WITH STREAMING DATA

Anjin Liu

Faculty of Engineering and Information Technology

University of Technology Sydney

A thesis submitted for the Degree of

Doctor of Philosophy

April 2018

CERTIFICATE OF AUTHORSHIP/ORIGINALITY

I certify that the work in this thesis has not previously been submitted for a degree nor has it been submitted as part of the requirements for a degree except as fully acknowledged within the text.

I also certify that the thesis has been written by me. Any help that I have received in my research work and the preparation of the thesis itself has been acknowledged. In addition, I certify that all information sources and literature used are indicated in the thesis.

Anjin Liu

April 2018

Acknowledgements

This has been a memorial and exciting journey. I would like to extend my warm gratitude to the people who inspired and helped me in many ways.

I would like to express my earnest thanks to my supervisors, Professor Guangquan Zhang and Distinguished Professor Jie Lu, for their knowledgeable suggestions and critical comments. During my doctoral research, their comprehensive guidance always illuminated the way. My discussions with them greatly improved the scientific aspect and quality of my research. Their strict academic attitude and respectful personality benefited my PhD study and will be a great memory throughout my life. I have learnt so much from them; it has been an honour.

I am honored to have met all the talented researchers of the Decision Systems & e-Service Intelligence Lab (DeSI). I have greatly enjoyed the pleasurable and plentiful research opportunities I shared with them. I would like to give my special thanks to Dr Ning Lu for whom inspired me to concept drift, and Feng Liu, Yiliao Song and Feng Gu with whom I engaged in concept drift-related research. The discussions with them were rewarding and fun.

I kindly thank Ms Jemima Moore and Ms Michele Mooney for polishing the language used in my thesis and publications. I have learnt much about academic writing from them.

I am grateful to the School of Software in the Faculty of Engineering and Information Technology at the University of Technology Sydney. This study was supported by the Australian Postgraduate Award (APA) and the Australian Research Council (ARC) discovery project.

Finally, I would like to express my heartfelt appreciation and gratitude to my parents and my wife for their love and support.

Abstract

The term concept drift refers to the change of distribution underlying the data. It is an inherent property of evolving data streams. Concept drift detection and adaptation has been considered an important component of learning under evolving data streams and has attracted increasing attention in recent years.

According to the existing literature, the most commonly used definition of concept drift is constrained to discrete feature space. The categorization of concept drift is complicated and has limited contribution to solving concept drift problems. As a result, there is a gap to uniformly describe concept drift for both discrete and continuous feature space, and to be a guideline to addressing the root causes of concept drift.

The objective of existing concept drift handling methods mainly focuses on identifying when is the best time to intercept training samples from data streams to construct the cleanest concept. Most only consider concept drift as a time-related distribution change, and are disinterested in the spatial information related to the drift. As a result, if a drift detection or adaptation method does not have spatial information regarding the drift regions, it can only update learning models or their training dataset in terms of time-related information, which may result in an incomplete model update or unnecessary training data reduction. In particular, if a

false alarm is raised, updating the entire training set is costly and may degrade the overall performance of the learners. For the same reason, any regional drifts, before becoming globally significant, will not trigger the adaptation process and will result in a delay in the drift detection process. These disadvantages limit the accuracy of machine learning under evolving data streams.

To better address concept drift problems, this thesis proposes a novel **R**egional **D**rift **A**daptation (RDA) framework that introduces spatial-related information into concept drift detection and adaptation. In other words, RDA-based algorithms consider both time-related and spatial information for concept drift handling (concept drift handling includes both drift detection and adaptation).

In this thesis, a formal definition of regional drift is given which has theoretically proved that any types of concept drift can be represented as a set of regional drifts. According to these findings, a series of regional drift-oriented drift adaptation algorithms have been developed, including the **N**earest **N**eighbor-based **D**ensity **V**ariation **I**dentification (NN-DVI) algorithm which focuses on improving concept drift detection accuracy, the **L**ocal **D**rift **D**egree-based **D**ensity **S**ynchronization **D**rift **A**daptation (LDD-DSDA) algorithm which focuses on boosting the performance of learners with concept drift adaptation, and the **online R**egional **D**rift **A**daptation (online-RDA) algorithm which incrementally solves concept drift problems quickly and with limited storage requirements. Finally, an extensive evaluation on various benchmarks, consisting of both synthetic and real-world data streams, was conducted. The competitive results underline the effectiveness of RDA in relation to concept drift handling.

To conclude, this thesis targets an urgent issue in modern machine learning research. The approach taken in the thesis of building regional concept drift detection

and adaptation system is novel. There has previously been no systematic study on handling concept drift from spatial prespective. The findings of this thesis contribute to both scientific research and practical applications.

Table of Contents

List of Figures

List of Tables