CONCEPT DRIFT DETECTION FOR MACHINE LEARNING WITH STREAM DATA

Feng Gu

Faculty of Engineering and Information Technology

University of Technology Sydney

A thesis submitted for the Degree of

Doctor of Philosophy

December 2019

CERTIFICATE OF ORIGINAL AUTHORSHIP

I, Feng Gu declare that this thesis, is submitted in fulfilment of the requirements for the award of doctorate, 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.

Feng Gu

December 2019

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 gratitude 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 am honored to have met all the talented researchers of the Decision Systems & e-Service Intelligence Lab (DeSI) at Centre for Artificial Intelligence. 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, Dr Anjin Liu and Dr Fan Dong who inspired me to concept drift research, as well as Feng Liu, Yiliao Song and Zhen Fang with whom I engaged in daily discussions that 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 Research Council (ARC) discovery project DP150101645.

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

Abstract

Machine learning in streaming data is often inhibited by arbitrary changes of the data distribution. Particularly, classification boundary change, also known as concept drift, is the major cause of machine learning performance deterioration.

Accurately and efficiently detecting concept drift remains challenging because of inherent limitations of stream data - non-stationarity, velocity and availability of true label data. The non-stationarity of the stream data causes performance degradation of pretrained models and the high velocity of the data generation requires highly efficient prediction algorithms for real time applications. The theoretical foundations of existing drift detection methods - two-sample distribution tests and monitoring classification error rate, both suffer from inherent limitations such as the inability to distinguish virtual drift (changes not affecting the classification boundary, will introduce unnecessary model maintenance), limited statistical power, or high computational cost. Furthermore, no existing detection method can provide information about the trend of the drift, which could be invaluable for model maintenance.

To better address concept drift problems, this thesis first proposes a novel concept drift detection method based on **N**eighbor **S**earch **D**iscrepancy (NSD), a new statistic that measures the classification boundary difference between two samples. The proposed method uses true label data to detect concept drift with high accuracy while

ignoring virtual drift. It can also indicate the direction of the classification boundary change by identifying invasion or retreat of a certain class, which is also an indicator of separability change between classes.

To improve concept drift adaptation efficiency, based on NSD, this thesis proposes two novel instance selection methods for both concept drift detection - **D**ecision **R**egion **S**upport Set (DRS) and classification - **D**ecision **R**egion **B**order Set (DRB). The unified framework yields reduction instances for both objectives simultaneously without computational overhead. The drift detection method efficiently detects concept drift without relying on resampling technique. The reduction rule based on Neighbor Search better estimates decision boundaries, resulting in improved classification accuracy.

For scenarios where true label data is unavailable, this thesis first proposes a novel distribution change detection method - **E**qual **D**ensity **E**stimation (EDE) based on the estimation of equal density regions. The aim is to overcome the issues of instability and inefficiency that underlie methods of predefined space partitioning schemes. This method is general, nonparametric and requires no prior knowledge of the data distribution.

Finally, in order to detect concept drift without true label data, this thesis introduces a novel categorization of drift types - maintainable and unmaintainable drift, to describe the necessity of model maintenance in different scenarios. Then we develop a unique drift detection algorithm based on **P**robability **P**ercentile **D**iscrepancy (PPD), which detects only maintainable drift without relying on true label data.

In summary, this thesis targets a critical issue in modern machine learning research. The approaches taken in the thesis of building effective and efficient concept drift detection algorithms are novel and practical. There has been no previous study on the theories of neighbor search discrepancy and maintainable concept drift. The findings of this thesis contribute to both scientific research and practical applications.

Table of Contents

List of Figures

List of Tables

List of Abbreviations