



**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 **Regional Drift Adaptation (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 **Nearest Neighbor-based Density Variation Identification (NN-DVI)** algorithm which focuses on improving concept drift detection accuracy, the **Local Drift Degree-based Density Synchronization Drift Adaptation (LDD-DSDA)** algorithm which focuses on boosting the performance of learners with concept drift adaptation, and the **online Regional Drift Adaptation (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 perspective. The findings of this thesis contribute to both scientific research and practical applications.

Table of Contents

CERTIFICATE OF AUTHORSHIP/ORIGINALITY	iii
Acknowledgements	v
Abstract	vii
List of Figures	xv
List of Tables	xvii
1 Introduction	1
1.1 Background	1
1.2 Research Questions and Objectives	4
1.3 Research Contributes	9
1.4 Research Significance	10
1.5 Thesis Structure	11
1.6 Publications Related to this Thesis	14
2 Literature Review	17
2.1 Concept Drift	17

2.1.1	Definition of concept drift and sources	17
2.1.2	Classification of concept drift	19
2.1.3	Related research topics and applications	21
2.1.3.1	Related research topics	21
2.1.3.2	Related applications	25
2.2	Concept Drift Detection and Adaptation	26
2.2.1	Concept drift detection	26
2.2.1.1	Drift detection framework	27
2.2.1.2	Concept drift detection algorithms	29
2.2.1.3	A summary of drift detection algorithms	37
2.2.2	Concept drift adaptation	39
2.2.2.1	Single learning model adaptation	39
2.2.2.2	Ensemble learning for concept drift adaptation	43
3	The Nature of Regional Drift	47
3.1	Introduction	47
3.2	Regional Drift and Regional Drift Presented Concept Drift	48
3.3	The Relationship between Regional Drift and Concept Drift	49
3.4	Summary	52
4	Concept Drift Detection via Accumulating Regional Density Discrepan-	53
	cies	
4.1	Introduction	53
4.2	Preliminary	55
4.3	Nearest Neighbor-based Data Embedding	57
4.3.1	Modelling data as a set of high-resolution partitions	59

4.3.2	Partition size optimization	67
4.4	Nearest Neighbor-based Density Variation Identification	69
4.4.1	A regional drift-oriented distance function	69
4.4.2	Statistical guarantee	71
4.4.2.1	Permutation test	71
4.4.2.2	A tailored significant test	74
4.4.3	Implementation of NN-DVI for learning under concept drift	77
4.5	Experiments and Evaluation	80
4.5.1	Evaluating the effectiveness of d^{mnp}	82
4.5.2	Evaluating the NN-DVI drift detection accuracy	89
4.5.3	Evaluating the NN-DVI on real-world datasets	98
4.5.4	Evaluating the stream learning with NN-DVI with different parameters	103
4.6	Summary	105
5	Concept Drift Adaptation via Regional Density Synchronization	107
5.1	Introduction	107
5.2	Local Drift Degree	110
5.2.1	The definition of LDD	110
5.2.2	The statistical property of LDD	110
5.3	Drifted Instances Selection and Adaptation	113
5.3.1	Drifted instance selection	113
5.3.2	Density synchronized drift adaptation	115
5.4	Experiment and Evaluation	118
5.4.1	Evaluation of LDD-DIS	118

5.4.2	Evaluation of LDD-DSDA	122
5.5	Summary	125
6	Incremental Regional Drift Adaptation	127
6.1	Introduction	127
6.2	A Regional Drift Adaptation Framework	129
6.3	Online Regional Drift Adaptation	131
6.3.1	kNN-based dynamic region construction	132
6.3.2	kNN-based regional drift detection	134
6.3.3	kNN-based regional drift adaptation	136
6.3.4	The implementation of online-RDA	136
6.4	Experiment and Evaluation	138
6.4.1	Evaluation of the capabilities of online-RDA on drift detec- tion and adaptation	138
6.4.2	Evaluation of online-RDA on synthetic drift datasets	140
6.4.3	Evaluation of online-RDS stream learning on real-world datasets	143
6.5	Summary	149
7	Conclusion and Future Research	155
7.1	Conclusions	155
7.2	Future Study	158
	Bibliography	161
	Appendix	185

List of Figures

1.1	A general framework of concept drift handling	3
1.2	A mapping from trends to challenges and research topics	4
1.3	Thesis structure	12
2.1	Three sources of concept drift	19
2.2	A demonstration of concept drift types	20
2.3	A general framework of concept drift detection	27
2.4	Landmark time window for drift detection	30
2.5	Two sliding time windows for drift detection	31
2.6	Two time windows for drift detection with fixed historical window .	33
2.7	Parallel multiple hypothesis test drift detection.	35
2.8	Hierarchical multiple hypothesis test drift detection.	36
3.1	Converting sudden drift and incremental drift to a set of regional drifts	51
4.1	Distribution-based drift detection framework	58
4.2	Conventional space partitioning methodology	60
4.3	Instance-oriented space partitioning	61
4.4	k-nearest neighbor-based instance-oriented space partitioning	62

4.5	Instance particle independence	68
4.6	A demonstration of accumulated regional density dissimilarity measurement	72
4.7	The test statistics of two-sample K-S test between normally distributed data with varying μ	84
4.8	The selection of k for NN-DVI	84
4.9	d^{mps} between normal distributed data (batch size 400) with varying μ	85
4.10	d^{mps} between normal distributed data (batch size 50) with varying μ	85
4.11	The test statistics for the two-sample K-S test between normal distributed data with a varying σ	86
4.12	The d^{mps} between normally distributed data that varies σ	87
4.13	1-D normal distribution with regional drift detection	88
4.14	NN-DVI classification accuracy of real-world datasets	104
5.1	A demonstration of the importance of considering regional drift	109
5.2	An illustration of how LDD works	111
5.3	LDD-DIS on Gaussian distribution with drifted variance	121
5.4	LDD-DIS on Gaussian mixture distribution with drifted Mean	121
6.1	A concept drift adaptation framework based on regional drift	130
6.2	Experiment evaluation of online-RDA on drift detection and adaptation	141
6.3	The average buffer size of online-RDA on synthetic datasets	145
6.4	The buffer size of online-RDA on real-world datasets	148

List of Tables

2.1	A summary of drift detection algorithms	38
4.1	NN-DVI drift detection results on $M(\Delta)$ stream	92
4.2	NN-DVI drift detection results on $C(\Delta)$ stream	93
4.3	NN-DVI drift detection results on $P(\Delta)$ stream	95
4.4	NN-DVI drift detection results on HD $C(\Delta)$ streams	97
4.5	NN-DVI average drift detection results	98
4.6	NN-DVI classification accuracy of real-world datasets	102
5.1	Comparison of LDD-DSDA and different data stream classification algorithms on real-world datasets	124
6.1	Online-RDA evaluation one-dimensional sudden-incremental drift data generator	139
6.2	Online-RDA evaluation synthetic data generator	142
6.3	Online-RDS evaluation the accuracy of synthetic datasets	144
6.4	Online-RDA evaluation real-world dataset characteristics	149
6.5	Online-RDA evaluation real-world datasets accuracy (%)	150
6.6	Online-RDA evaluation real-world datasets execution time (ms)	151

6.7 Online-RDA evaluation real-world datasets memory cost (GB RAM-
Hours) 152