UNIVERSITY OF TECHNOLOGY SYDNEY

Handling Concept Drift Using the Correlation between Multiple Data Streams

by Bin Zhang

Thesis submitted in fulfilment of the requirements for the degree of

Doctor of Philosophy

under the supervision of Dist. Prof. Jie Lu and A/Prof. Guangquan Zhang

University of Technology Sydney
Faculty of Engineering and Information Technology

October 2022

CERTIFICATE OF ORIGINAL AUTHORSHIP

I, *Bin Zhang*, declare that this thesis is submitted in fulfillment of the requirements for the award of *Doctor of Philosophy*, in the *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 Research Council.

Production Note:

Signature: Signature removed prior to publication.

Date: October 24th, 2022

ABSTRACT

Machine learning has been widely applied to handle big data. In real-world applications, data are in the form of streams. Streaming data bring new challenges to machine learning. Concept drift is a major problem in handling streaming data. The newly arrived data have a different distribution from historical data. Hence, machine learning models do not work on the newly arrived data. Many methods have been proposed to detect whether concept drift occurs and to update the machine learning models to the drift.

To date, the research on concept drift considers data streams separately. That is, these methods handle data streams one by one, ignoring whether correlations exist between data streams. However, in real-world applications, correlations between data streams are widespread. If the correlations between data streams were available, this would support better decision making, rather than handling data streams separately. There is currently no research on how to model the correlations between data streams. Data streams are dynamic. Therefore, it is necessary to develop methods to track the correlations between data streams and to adapt to changes in correlations. In addition, after concept drift is detected, there is usually insufficient data to train a new model, which can lead to the over-fitting problem. How to deal with the over-fitting issue is still a challenge. Motivated by this, this research proposes several methods to overcome the aforementioned challenges.

To alleviate the over-fitting problem, a concept drift adaptation method, Drift Adaptation via Joint Distribution Alignment (DAJDA), is proposed. DAJDA performs a linear transformation to the drift instances instead of modifying the model. Instances are transformed into a common feature space, reducing the discrepancy of distributions before and after drift. Using additional historical data to train a new model can lead to better performance due to the increasing training set. Experimental studies show that DAJDA has the ability to improve the performance of the learning model under concept drift.

To model the correlations between data streams, we propose a novel Multi-stream Concept Drift Handling Framework via data sharing, containing fuzzy membership-based drift detection (FMDD) and fuzzy membership-based drift adaptation (FMDA) components, to train the new learning model for drifting streams by sharing weighted data from other non-drifting streams. A stream fuzzy set is defined with membership functions that measure the degree to which samples belong to a data stream. Our Concept Drift Handling Framework can detect when and in which streams concept drift occurs, and therefore, the over-fitting issue can be solved by adding the weighted data from non-drifting streams to train new learning models. Synthetic and real-world experiment results show that our method can help avoid the over-fitting issue caused by a lack of data and thereby significantly improve the prediction performance.

To track changes in correlations and adapt to correlation drift, we propose an ensemble chain-structured model, Evolutionary Regressor Chains. To provide the model with the ability to search the optimal order of the chain, we design a heuristic order searching strategy to be incorporated as part of the model's ongoing process. The heuristic order searching strategy can also update the chains as time passes, to track the dynamicity of the correlations. A diversity pruning method is also proposed to reduce computation complexity while retaining the diversity of the ensemble. We undertake a theoretical analysis, and give the dynamic regret bound of our method. Our experiment results show that our Evolutionary Regressor Chains method can track data stream correlations accurately. Chains can update themselves to adapt to both concept drift and correlation drift. The performance of the machine learning models on data streams is improved.

To learn meta knowledge across multiple data streams, a concept drift adaptation strategy - Learning to Fast Adapt in the Evolving Environment - is proposed for neural network classifiers in the non-stationary environment. A meta-level LSTM recurrent neural network is used to learn a proper parameter updating rule instead of updating methods that are traditional gradient descent-based (e.g. stochastic

gradient descent) during fine-tuning. A suitable updating step will be generated according to current loss and its gradient for the parameter of a neural network classifier. Experiments on both synthetic and real-world data sets show that our method can quickly adapt the neural network classifier to concept drift and help improve the performance of the classifier in a dynamic environment.

Dedication

I dedicate my dissertation work to my parents Xia Xiang and Xuxin Zhang.

Acknowledgements

It has been an indelible journey to complete my doctoral degree at the esteemed University of Technology Sydney. I am profoundly grateful to all those who have extended their assistance in myriad ways throughout this remarkable endeavor.

Foremost, I wish to express my sincerest appreciation to my principal supervisor, the highly esteemed Distinguished Professor Jie Lu. Her unwavering guidance has been an illuminating beacon, navigating me towards the successful culmination of my research. In addition to fostering an exceptional research environment, she bestowed upon me a generous scholarship that served as a solid bedrock for my four-year inquiry. Under her tutelage, I delved into uncharted research domains, identifying and addressing critical gaps with her expert counsel. Her unwavering dedication to scholarship served as a constant source of inspiration and solace during challenging times. Beyond research, her invaluable professional advice on life and potential career trajectories has shaped my path with utmost clarity. Moreover, her pursuit of innovation and unwavering commitment to scholarly rigor has left an indelible mark on my academic journey.

I also extend my deepest gratitude to my esteemed co-supervisor, Associate Professor Guangquan Zhang. His unwavering support and invaluable suggestions pertaining to literature review, manuscript writing, and paper revision have been instrumental in shaping the trajectory of my doctoral study. It is his trust, encouragement, and unwavering assistance that have played an integral role in my successful completion of this demanding program.

Joining the Decision Systems & e-Service Intelligence Lab has been an absolute pleasure, providing me with an enriching experience alongside numerous enthusiastic and dedicated colleagues. I wish to express my heartfelt gratitude to Dr. Anjin

Liu, Dr. Yiliao Song, Dr. Hang Yu, Dr. Bin Wang, and my fellow peers who have generously offered their guidance and suggestions throughout the course of my doctoral candidature.

Additionally, I would like to extend my sincerest appreciation to my parents, whose unwavering support has been a constant source of strength regardless of my location or pursuits. I am forever indebted to their encouragement and unwavering belief in my abilities. To my dear friends Donglai Yang and Shiyu Ma, I cherish the memories of our shared experiences and remain deeply grateful for your unwavering companionship during the challenging times imposed by the Covid-19 pandemic. I would like to express special appreciation to my best friend, Zihe Liu, whose presence has brought immeasurable joy and laughter to my life. Our bond is truly extraordinary, and I am immensely thankful for the camaraderie we share.

In summary, the successful completion of my doctoral journey would not have been possible without the support, guidance, and unwavering belief of those mentioned above. Their collective contributions have left an indelible mark on my academic and personal growth, and for that, I am eternally grateful.

Bin Zhang

Sydney, Australia, 2022.

Contents

List of Figures

List of Tables

Abbreviation

ARF: Adaptive Random Forest

DAJDA: Drift Adaptation via Joint Distribution Alignment

DDM: Drift Detection Method

ECDD: EWMA for Concept Drift Detection

ERC: Ensemble Regressor Chains

FMDA: Fuzzy Membership-based Drift Adaptation

FMDD: Fuzzy Membership-based Drift Detection

HAT: Heoffding Adaptive Tree

HDDM: Hoeffding's inequality based Drift Detection Method

HT: Heoffding Tree

JDA: Joint Distribution Adaptation

MMD: Maximum Mean Discrepancy

MOA: Massive Online Analysis

MTRS: Multi-target Regression Stacking

NSW: New South Wales

PAC: Probably approximately Correct learning

PCA: Principal Component Analysis

RC: Regressor Chains

SEA: SEA Moving Hyperplane Concepts

ST: Single Target

VIC: Victoria

Nomenclature and Notation

$X_t^{(i)}$ is the feature at time t of the ith stream.

$y_t^{(i)}$ is the label at time t of the ith stream.

S_i is the ith stream.

$P(\cdot)$ is the probability.

$E(\cdot)$ is the expect.