UNIVERSITI TEKNOLOGI MARA

ENHANCING HIGH-DIMENSIONAL STREAMING DATA ANALYSIS: OPTIMIZING ONLINE FEATURE SELECTION FOR HANDLING DRIFT USING OPTIMIZATION TECHNIQUE AND ENSEMBLE LEARNING

EZZATUL AKMAL BINTI KAMARU-ZAMAN

Thesis submitted in fullfillment
of the requirements for the degree of
Doctor of Philosophy
(Computer Science)

College of Computing, Informatics and Mathematics

June 2024

ABSTRACT

In the era of data-driven decision-making, managing dynamic data streams characterized by evolving data distributions and high dimensionality presents a formidable challenge for online feature selection. This research addresses the challenge by developing innovative solutions in optimizing Online Feature Selection (OFS) to manage feature irrelevancy and redundancy, tackling the issues of Feature Drift, and rigorously validating the proposed algorithms in high-dimensional dynamic data streams. The research employs a structured methodology, introducing two novel methods: PSO-OSFS (Particle Swarm Optimization for Online Streaming Feature Selection), an optimized online feature selection and its enhancement, PSO-OSFS+ HEFT designed to handle feature drift. The PSO-OSFS method is underpinned by the adaptive threshold particle representation of particle swarm optimization and enhanced fitness function using minimization of mean absolute deviation of dependency among feature subsets. Adaptive threshold particle representation introduces a novel aspect in defining a threshold value of significance level from 0.01 to 0.1. This unique contribution sets the research apart in the field. PSO-OSFS+HEFT combines windowing as a drift adaptation method with the hybridization of PSO-OSFS and a benchmark ensemble learning approach named Heterogeneous Ensemble Method for Feature Drift (HEFT). Compared to original HEFT that uses traditional feature selection, the proposed method is able to mitigate the issue of batch learning of online feature selection while handling feature drift. The experiment phase involves highly redundant and relevant datasets, various window sizes, and different feature drift types. PSO-OSFS emerges with an impressive average of 76.8% accuracy, outperforming its counterparts. Meanwhile, PSO-OSFS+HEFT consistently achieves a remarkably high accuracy of average 69.30%, with 1.82% higher than other method and able to peak at 91.28%. Further refinement strategies, including noise injection and rebalancing elevate the proposed method's robustness, leading to better average accuracy of 70.47%, with increment of 1.17% from without refinement. Comparison with the original HEFT reveals that the proposed PSO-OSFS+HEFT consistently opts for a lower feature subset selection with a non-significant accuracy dip at a minimal 0.01 difference. Validation reinforces the insight that the proposed method adeptly strikes an optimal equilibrium between feature reduction and model performance. The higher median feature importance values with fewer outliers, affirm the suitability of the proposed method for real-world applications where data evolves such as cybersecurity, finance, healthcare and more. In conclusion, the results demonstrated the significant contributions of the method in enhancing model accuracy, adapting to evolving data distributions, optimizing feature subsets, and bolstering model resilience in a dynamic data stream environment. The research is expected to advance the field of data science and empowering end-users to make informed decisions under dynamic data stream circumstances.

ACKNOWLEDGEMENT

I am extremely grateful to God for allowing me to successfully complete my PhD after a long and challenging journey. First and foremost, I would like to express my heartfelt thanks to my supervisors, Dr. Azlin Ahmad and Prof. Dr. Hajah Azlinah Haji Mohamed. Their unwavering guidance, remarkable knowledge, and endless patience have inspired me to become the best version of myself.

I am also grateful to the Ministry of Higher Education Malaysia and Universiti Teknologi MARA, UiTM for providing funding for this study. Also, I would like to extend my appreciation to the Assistant Vice Chancellor, Professor Ts. Dr Haryani Haron; Head of Postgraduate, PM Dr Norjansalika Janom; School of Computing Sciences Head,PM Dr Mohamad Yusof Darus; and Intelligent Systems Department Coordinator, Dr Shamimi Abdul Halim. I am filled with gratitude for the examiners who have provided me with uplifting comments during my DRP and Mock VIVA sessions. Their words of encouragement have inspired me to push forward and strive for greatness.

I owe my success to the unwavering support of my loved ones. My parents, Haji Kamaru-zaman Muhammad and have instilled in me the importance of education and provided me with the foundation to achieve my dreams. I am also grateful to my siblings, in-laws, and friends for their constant encouragement and support. Special thanks to my PhD buddy, Puan Nur Huda Nabihan Md Shahri, for always checking up on me and motivating me. Most importantly, to my husband and children, Haji Muhammad Syahmi Jamaluddin, M. Arsyad Ziqri, Aufa Humaira, and Ayesha Hana, thank you so much for being my unwavering sources of support and my pillars of strength. This victory is dedicated to all of them who believed in me and inspired me to keep going.

TABLE OF CONTENTS

CHAPTER 1
INTRODUCTION

In today's data-driven landscape, the generation and flow of information have reached unprecedented levels. Streaming data, characterized by its continuous, real-time nature, has become ubiquitous in numerous domains. The dynamic nature of streaming data introduces a multitude of complexities, chief among them being the need for effective feature selection and the management of feature drift. This chapter provides details about the research background on feature selection, leading to the several issues that arise from feature selection. Critical issues of feature selection when handling streaming data will be discovered significantly. Online feature selection (OFS) will be generally explained in the research background as an approach to handle the issue of streaming data. However, the ever-evolving characteristics of data streams render conventional feature selection approaches insufficient. This chapter delves into the intricate landscape of dynamic data streams, where traditional batch learning and OFS methods fall short. It explores the underlying problems, intricacies, and challenges of OFS and feature drift in dynamic data environments. The issues and challenges of OFS and feature drift will be related to the performance measure to investigate the effect of different data dimensionalities. These issues and challenges will support the problem statement's elaboration and lead to the research objectives. In summary, this chapter focused on the research background, problem statement, research questions, research objective, research scope, research significance and novel contribution of the proposed study.

1.1 Research Background

The unprecedented scale of data that emerged from different sources, such as the Internet of Things (IoT), social media, bioinformatics, and other data resources, significantly impacted the Big Data research area (Mohamed et al., 2020). The emergence of 7V's of Big Data aspects such as variety, velocity, volume, veracity, variability, visualization, and value (Kapil et al., 2016; Thudumu et al., 2020) demands extensive research toward knowledge discovery and pattern. These aspects have war-