Chapter 1
Unveiling the Role of Diversity Measure in Concept Drift Detection: A Comparative Analysis and Future Prospects

Osamah A. Mahdi
Melbourne Institute of Technology, Australia

Nawfal Ali
Monash University, Australia

Eric Pardede
La Trobe University, Australia

Bhagwan Das
https://orcid.org/0000-0002-0731-6353
Melbourne Institute of Technology, Australia

ABSTRACT

Concept drift, marked by changes in the statistical properties of a target variable, is a notable challenge in machine learning, data mining and applications involving big data and large-scale data processing. Addressing this, the employment of diversity measure has emerged as an effective strategy. This chapter examine and investigate the role of the diversity measure in detecting concept drift and compare and analysis four different ways of using them: DMDDM for drift` detection in a fully supervised binary classification context, DMDDM-S in a semi-supervised context, DMODD for online drift detection in a fully supervised multi-classification context,

DOI: 10.4018/979-8-3693-5448-3.ch001

and HBBE, a hybrid block-based ensemble designed for addressing different types of concept drifts. The comparative analysis evaluates the efficacy of these methods and the results confirm their effectiveness within their respective settings. Overall, this chapter explores advancements in using diversity measures for concept drift detection in big data, emphasizing their importance and future research in machine learning contexts.

INTRODUCTION

The constant transformation or evolution of data is a crucial concern in dynamic settings and applications, including aviation, autonomous vehicles, nuclear reactors, healthcare, the military, smart cities, and the aerospace sector. Essentially, the critical characteristics of these environments are subject to change, potentially leading to negative consequences, such as endangering human lives, if not adequately addressed (Gama 2010). Consequently, learning methods must employ sophisticated algorithms to monitor these changes and adapt accordingly. Furthermore, the effectiveness of learning algorithms may vary due to the changing nature of incoming data, meaning that an algorithm that is effective today may become outdated following changes in the environment or data.

The literature on learning from data streams identifies the phenomenon of class distribution changes within data streams as concept drift (Mahdi, Ali, Pardede, Alazab, et al. 2024). In the context of machine learning, concept drift describes a situation where the statistical properties of the target variable, which the model is designed to forecast, shift over time (Prasad and Agarwal 2016). This implies that the original input data's relevance to the model has altered significantly, yet the model remains oblivious to these modifications and, as a result, fails to make precise predictions. Thus, it is imperative for learning algorithms to detect concept drift in dynamic data streams and accordingly adjust or renew their prediction models. To address this challenge, adaptive learning models are developed, employing drift detection methods to pinpoint the instances of drift in changing environments (Mahdi, Ali, Pardede, Alazab, et al. 2024).

The rapid alterations in human behavior triggered by the COVID-19 pandemic serve as a stark illustration of concept drift. To explore the impact of pandemic-induced concept drift on predictive models, consider Melbourne City's scenario, where quarantine measures have influenced various aspects of daily life, including shopping habits, electricity consumption, and traffic accidents. For instance, analyzing predictions for future traffic accidents, VicRoads, which documents all incidents involving injuries, fatalities, or property damage exceeding $1,000, noted a marked decrease in annual accident figures during the pandemic due to lockdowns and the

Evolving Business Intelligence on Data Integration, ETL Procedures, and the Power of Predictive Analytics