IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 21 October 2022, accepted 15 November 2022, date of publication 28 November 2022,
date of current version 8 December 2022.

Digital Object Identifier 10.1109/ACCESS.2022.3225225

RESEARCH ARTICLE

Adaptive Learning With Extreme Verification Latency in Non-Stationary Environments

MOBIN M. IDREES[1], FREDERIC STAHL[2], AND ATTA BADII[1]
[1]Department of Computer Science, University of Reading, RG6 6AH Reading, U.K.
[2]German Research Center for Artificial Intelligence GmbH (DFKI), Marine Perception, 26129 Oldenburg, Germany

Corresponding author: Frederic Stahl (frederic_theodor.stahl@dfki.de)

This work was supported by the Ministry for Science and Culture, Lower Saxony, Germany, through Funds from the Niedersachsische Vorab, under Grant ZN3683 and Grant ZN3480.

ABSTRACT Existing Data Stream Mining algorithms assume the availability of labelled and balanced data streams. However, in many real-world applications such as Robotics, Weather Monitoring, Fraud-Detection systems, Cyber Security, and Human Activity Recognition, a vast amount of high-speed data is generated by Internet of Things sensors and real-time data on the Internet are unlabelled. Furthermore, the prediction models need to learn in Non-Stationary Environments due to evolving concepts. Manual labelling of these data streams is not practical due to the need for domain expertise and the time-resource-prohibitive nature of the required effort. To deal with such scenarios, existing approaches are self-Learning or Cluster-Guided Classification (CGC) which predict the pseudo-labels, which further update the prediction models. Previous studies have yet to establish a clear and conclusive view as to when, and why one pseudo-labelling approach should be preferable to another and what causes an approach to fail. In this research, we propose a novel approach, "*Predictor for Streaming Data with Scarce Labels*" (PSDSL), which is capable of intelligently switching between self-learning, CGC and micro-clustering strategies, based on the problem it is applied to, i.e., the different characteristics of the data streams. In PSDSL a novel approach called *Envelope-Clustering* has been introduced to resolve the conflict during the cluster labelling which suggested a confidence measure approach to ensure the quality and correctness of labels assigned to the clusters. The auto parameter tuning mechanism of PSDSL eliminates the human dependency and determines the best value of number of centroids from initial labelled data. The predictive performance of the PSDSL is evaluated on non-stationary datasets, synthetic data-streams, and real-world datasets. The approach has shown promising results on randomised datasets as well as on synthetic data-streams, as compared with state-of-the-art approaches. This is the first large-scale study on an adaptive extreme verification approach that supports automatic parameter tuning and intelligent switching of pseudo-labelling strategy, thus reducing the dependency of machine learning on human input.

INDEX TERMS Concept drift, data stream mining, extreme verification latency, non-stationary environment, semi-supervised learning.

I. INTRODUCTION

Data Stream Mining algorithms assume the availability of labelled data, immediately or after some delay, to update the accuracy of the classifier and update the prediction model. However, with certain applications such as Fraud-Detection Systems [1], Cyber Security [2], [3] and Human Activity Recognition [4], [71] etc. the data stream is unlabelled and manual labelling is impractical due to the cost, time, and the need for domain expertise. In machine learning literature, this scenario is referred to as Verification Latency (VL) [5], [6]. In another scenario, only limited labelled data is followed by completely unlabelled instances; this scenario is referred to as Extreme Verification Latency (EVL) [7], [8], [10], [11].

The associate editor coordinating the review of this manuscript and approving it for publication was Zhan Bu.