Cluster Computing (2022) 25:4481–4494
https://doi.org/10.1007/s10586-022-03695-z

# Adaptive PCA-based feature drift detection using statistical measure

Supriya Agrahari[1] · Anil Kumar Singh[1]

Received: 9 March 2022 / Revised: 9 March 2022 / Accepted: 19 July 2022 / Published online: 5 August 2022

## Abstract

The plethora of existing methods in the streaming environment is sensitive to extensive and high-dimensional data. The distribution of these streaming data may change concerning time, known as concept drift. Several drift detectors are built to identify the drift near its occurrence point. Still, they lack proper attention to determine the feature relevance change over time, known as feature drift. Over time, the distribution change of the relevant features subset or the change in the relevant features subset itself may cause feature drift in the data stream. The paper proposes an adaptive principal component analysis based feature drift detection method (PCA-FDD) using the statistical measure to determine the feature drift. The proposed work presents a framework for identifying the most important features subset, feature drift, and incremental adaptation of the prediction model. The proposed method finds the relevant features subset by utilizing the incremental PCA and detects feature drift by observing the change in the percentage similarities among the most important features subset with respect to time. It also helps to forecast the prediction error of the base learning model. The proposed method is compared with state-of-the-art methods using synthetic and real-time datasets. The evaluation results exhibit that the proposed work performs better than the existing compared methods in terms of classification accuracy.

**Keywords** Data stream · Principal component analysis (PCA) · Feature drift · Concept drift · Prediction model

## 1 Introduction

With the advancement of technologies, many information society fields generate vast amounts of streaming data such as network access logs, weather forecasting data, medical data, traffic monitoring data, etc. The data stream is a set of data observations that arrives sequentially instance by instance. The traditional machine learning methods are based on the assumption of stationary data distribution. It means that the data is collected before the learning process Whereas the streaming environment contains the non-stationary distribution of data. Traditional machine learning methods fail to handle continuously generating data. Thus, various methods have been developed to solve data stream problems. The change in the distribution of data instance happens due to the dynamic nature of data over a period of time, known as concept drift [1].

There are different types of concept drifts like sudden or abrupt, incremental, gradual, recurring, and blip occurred in streaming data. These drifts are based on the speed of change in the data distribution concerning time. The concept drift negatively affects streaming data analysis and forecasting. In several drift detection methods, the detectors and learning model run simultaneously [2]. The detectors detect the distribution change in the incoming data instances, and the outcomes (or target values) of incoming data instances are predicted by the learning (or prediction) model. The learning model executes independent of the drift detection method, and its benefit is that it gives information about the dynamics of the generated data [3]. Due to the drift, it is seen that the decision boundary changes between the target (or class) values. Thus, the drift detection problem becomes more challenging because the training of the current prediction model is based on the old decision boundary. In this case, it wrongly classifies data instances (or examples) in the old class, whereas they generate from the new class. Several existing methods are based on monitoring the prediction error of the base

✉ Supriya Agrahari
supriyagrahari@gmail.com; supriyaagrahari@mnnit.ac.in

Anil Kumar Singh
ak@mnnit.ac.in

[1] MNNIT Allahabad, Prayagraj, India