Date of publication DEC-31, 2022, date of current version OCT-27, 2022.
www.computingonline.net / computing@computingonline.net

Print ISSN 1727-6209
Online ISSN 2312-5381
DOI 10.47839/ijc.21.4.2778

An Enhanced Online Boosting Ensemble Classification Technique to Deal with Data Drift

RUCHA C. SAMANT, SUHAS PATIL
Computer Engineering Department, College of Engineering, Bharati Vidyapeeth Deemed to be university, Pune, Maharashtra, India

Corresponding author: Rucha C. Samant (e-mail: ruchasamant25@gmail.com).

ABSTRACT Over the last two decades, big data analytics has become a requirement in the research industry. Stream data mining is essential in many areas because data is generated in the form of streams in a wide variety of online applications. Along with the size and speed of the data stream, concept drift is a difficult issue to handle. This paper proposes an Enhanced Boosting-like Online Learning Ensemble Method based on a heuristic modification to the Boosting-like Online Learning Ensemble (BOLE). This algorithm has been improved by implementing a data instance that retains the previous state policy. During the boosting phase of this modified algorithm, the selection and voting strategy for an instance is advanced. Extensive experimental results on a variety of real-world and synthetic datasets show that the proposed method adequately addresses the drift detection problem. It has outperformed several state-of-the-art boosting-based ensembles dedicated to data stream mining (statistically). The proposed method improved overall accuracy by 1.30 percent to 14.45 percent when compared to other boosting-based ensembles on concept drifted datasets.

KEYWORDS Boosting; concept drift; drift detectors; data stream mining; ensemble classification.

I. INTRODUCTION

Normally the goal of data analysis is to uncover hidden information. In many cases, this knowledge is useful for improving the working model or avoiding hazardous situations. The meaning of data has changed significantly over the last few decades as the sources of data generation have shifted from static to dynamic. The rate of data generation has increased exponentially as a result of the widespread use of online applications and devices, and the nature of data has also changed. Data is now referred to as a data stream, and one of its most significant characteristics is time. As a result, it continues to change over time, becoming increasingly critical in assessing.

Since this data is in the form of streams, it has unique properties that make interpreting and using various machine learning techniques inappropriate. The following are the main data stream processing challenges that need to be resolved, according to investigations [1, 2] on data stream mining literature:

1. The speed with which data streams are created implies that processing the complete data set may be delayed, or may be impossible.

2. Since the size of the stream is so large, it requires more memory than simple applications.

3. Concept drift happens when the underlying data distribution changes in the stream. Because the data is not static, the analysis of each and every example across time may alter. As a result, each slot's prediction changes with time, and the model constructed for such data must be versatile.

Classification is the foundation for analysis in most applications, such as market trends based on customer demands, spam mail rectification, climate change, and share market, to name a few [3]. However, an adaptable algorithm is needed for classifying data streams because of the enormous speed and amount of the data as well as the changing behavior of the data. Traditional single classifiers such as SVM [4] and Decision trees [5] are not sufficient to deal with all of these challenges. Advanced methods, such as adaptive sliding windows, data set sampling algorithms, drift detectors, and adaptive ensembles, have been developed in the literature to face this issue [6–14]. Ensemble classifiers outperformed all of these methods. Similarly, to process huge amount of data, more memory is required. Despite these drawbacks, ensemble has demonstrated its ability to deal with vivid data, has achieved high accuracy, and its simple design methods have drawn the attention of many researchers[1, 2, 15].

Either the short-term or long-term behaviour of a growing data stream may be more relevant in the categorization process