INCREMENTAL AGGREGATION MODEL FOR DATA STREAM CLASSIFICATION

S. Jayanthi[1] and B. Karthikeyan[2]
[1]Department of Computer Science and Engineering, Karpagam University, Coimbatore, India
[2]Dhanalakshmi Srinivsan Institute of Research and Technology, Siruvachur, Perambalur, India
E-Mail: nigilakash@gmail.com

ABSTRACT

In online data stream processing, data stream classification task confronts several challenges such as, concept drift, concept evolution and partial labeling due to the dynamic nature of data streams. Amid these issues, concept drift is on the top concern that degrades the accuracy of data stream classification task, immediately upon its occurrence. However, concept evolution and partial labeling are also equally notable plights that are not focused by most of the existing approaches. Ensemble learning is a widely accepted prominent method that attempts to reconcile the issues encountering in the data stream classification. Our previous work addresses only the different types of concept drifts. This paper expounds a Novel Incremental Aggregation Model (IAM) which makes use of Adaptive Probabilistic Neural Network (APNN), Aggregate Weighted Ensemble Model (AWEM) and Ensemble Cloning that makes the system impeccable by combating against all the above said issues. The performance of the proposed algorithm has been experimentally tested with few synthetic data sets. Experimental results show that our model outperforms the existing ensemble approaches in terms of accuracy.

Keywords: data stream processing, data stream classification, concept drift, concept evolution, partial labelling, ensemble learning;

1. INTRODUCTION

In this technological era, streams of incremental data are being generated in almost all digitalized organizations. Extracting knowledge from such data streams is the key for the success of these organizations. In general, data streams are massive in size, dynamic in nature, infinite in length. Hence, conventional data mining techniques which work well only on stationary data become unsuitable for handing dynamic data streams. Moreover, Data stream processing techniques have several resource constraints such as, single scanning of data streams, limited memory size and processing time. Indeed, these constraints might not be met with conventional classification techniques [12], [15], [16], [17].

Among several task of data stream processing, data stream classification is a most prominent supervised task that predicts and classifies the upcoming data streams in ever-changing data distribution center [21]. While data stream processing, data stream classification task confronts several challenges, such as, concept drift, concept evolution, partial labeling, outlier and real time analysis. Concept drift is one of the most common phenomenons in data stream processing that occurs due to any of the changes in data distribution center, interest besides the target concept and the rules underlying the classification task [2], [8], [11]. In general, incremental learning data set is subject to concept drift. Due to this, almost all incremental learning algorithms have constant look out on concept drift. Concept evolution and partial labeling occurs due to the emergence of novel classes and unlabelled instances respectively.

This paper is segmented into six sections. Section two discusses some of the research issues stood behind the data stream classification task. Section three addresses some of the most cited related work of the proposed approach. Section four expounds the proposed novel incremental aggregation model and its architecture. Section five illustrates the experimental results and the last section is concluded by instilling the tactics for the enhancement of the proposed work.

2. RESEARCH ISSUES

The following research issues are the motivation behind the proposed research work.

A. It is found that many of the data stream classification algorithms in the literature do not forecast about concept drift and works well only with stable data distribution centre.
B. Few algorithms are good in confronting different types of concept drifts altogether. That is the algorithm that confronts gradual concept drift efficiently fails to be good in other type of concept drifts, and vice versa.
C. Even the methods good in handling concept drift are not good in other issues such as novel class occurrence and partial labeling. For example, our previous model named Aggregate Weighted Ensemble Model exclusively focused on combating against different types of concept drifts. However, this model did not forecast about the concept evolution and partial labeling [22], [23].
D. Few approaches attempted to expound the method for handling novel class occurrence and partial labeling [23].
E. Finally, no one method is good in confronting all the above said issues altogether in dynamic data streams.
F. In light of these challenges, the proposed work is focused on defending against all the above said issues.