A New Approach for Detecting and Handling Concept Drift in Big Data in Order to Perform Accurate Classification

By:

Hisham Abdalmalik Ogbah

Supervised By:

Dr. Abdallah Alashqur

This Thesis was submitted in Partial Fulfillment of the Requirements for the Master's Degree In Computer Science

Applied Science University

Deanship of Scientific Research and Graduate Studies

Amman - Jordan

June 2017

A New Approach for Detecting and Handling Concept Drift in Big Data in Order to Perform Accurate Classification

By: Hisham Abdalmalik Ogbah

Supervised by: Dr. Abdallah Alashqur

Abstract

In classification, data tuples are mapped to a limited number of classes. The classifier learns a classification model from a pre-classified dataset. The classifier can then predict the class of newly added data based on the model that it learned. After the model is built, a concept drift may occur in the data due to changes in style or trend. The concept drift results in a new distribution of the underlying data relative to their classes. The problem of existing systems is that detecting the concept drift in large dataset is not efficient. Moreover, the handling of the drift does not take the drift intensity into consideration while generating a new model.

To solve these problems, this thesis introduces a new approach for coping up with concept drift, which consists of three parts. First, it introduces a new efficient algorithm for detecting the occurrence of a concept drift that is based on the idea of binary search, which is more efficient for large datasets. Second, a new way of measuring the intensity of the drift is introduced. Measuring the intensity of the drift is important because it impacts how we may choose to deal with it going forward. Third, a new algorithm is introduced for handling concept drift. What distinguishes this algorithm from other existing algorithms is that it uses a weighting technique based on the intensity of the drift. If the drift intensity is high, the model generation process discards old data and builds a new model solely based on the new data after drift. On the other hand, if the drift intensity is medium or low, the model generation

process takes into account both old data and new data but it gives more weight (proportional to the drift intensity) to the new data as compared to old data. This can be particularly useful when the drift intensity is not very severe and consequently the older data still has some merit associated with it.

Finally, a performance comparison with another robust detection algorithm shows that the new detection algorithm exhibits superior performance in terms of finding the drift position especially as the size of dataset is increased. On the other hand, the experiments conducted to evaluate the effectiveness of the handling algorithm that adapts itself based on the drift intensity shows good results and it can be relied on to build an accurate classification model.

Keywords*:* Big Data, Data Mining, Classification, Concept Drift, Drift Detecting, Drift Handling.