On-The-Fly Dynamic Classification of Evolving Text Data-Streams in a Neural Network-Based Optimum Ensemble Framework*

Evrilen Metin Veri-Akışlarının Dinamik Olarak Sinir Ağı Tabanlı Optimum Heyet Yapısıyla Anında Sınıflandırılması

TÜBİTAK Project No. 120E103

Fazlı Can (PI), Seyit Koçberber (I)

July 15, 2020

Topic. In today's world, a large number of Internet sources very quickly produce a numerous and very wide variety of limitless data. On-the-fly classification of data stream items is valuable to relevant systems, organizations and individuals for monitoring developments. Change in the relationship between data definitions and classes in evolving data streams is referred to as concept drift. In this study, we aim to develop a solution to the multi-class classification problem, which involves the selection of only one class in evolving data streams, by using artificial neural networks (neural networks, for short) and an ensemble classifier.

Contributions. In this project we investigate four problems that will be studied for the first time in this work. They are 1) The mathematical derivation and GPU-based development of two online neural network classifiers for text data streams; 2) The use of these classifiers as members of an geometrically optimum ensemble structure that minimizes the distance among the predictions of ensemble members and correct classes in the decision space; 3) Investigating if the highest prediction accuracy is observed, in accordance with the theoretical deduction, when the number of member neural networks of an ensemble is equal to the number of classes; and 4) Adapting a new concept drift detection algorithm to a neural network-based optimal ensemble structure so that the ensemble effectively evolves to new developments.

Method. We plan to adapt two newly proposed neural network architectures to the online classification domain and use them as members of an ensemble classifier. The first neural network classifier we plan to adapt, the Broad Learning System (BLS), has a single layer that can grow vertically. This single layer structure enables it to considerably limit the training time. The other structure, the Hedge Backpropagation (HBP), is composed of multiple layers whose count changes dynamically to solve the time problem that can be caused by backpropagation. The proposed neural network structures will be developed on a GPU environment and be used

as the members of an ensemble algorithm, GOOWE, which has been recently developed by our research group to make geometrically optimum decisions.

Project Management. Consists of four stages: 1. Selection of datasets with a wide variety and preparing a test collection for Turkish and structuring them in accordance with our objectives; 2. Mathematical derivation and GPU-based implementation of the chosen neural network approaches, and preparing the optimal GOOWE ensemble environment within the framework of neural networks; 3. Preparation of baseline algorithms; and 4. Comparative evaluation of approaches with different datasets and different concept drift detection scenarios with statistical tests.

Impact. The neural network structures and algorithms we will develop can be used for information filtering and information retrieval in news and social media platforms. They can be modified to suit new event detection and monitoring applications. The new concept drift detection algorithm, which we will adapt to the text data stream environment, can be generalized to the data stream clustering problem. The large-scale Turkish dataset we will prepare will encourage and support various text classification studies in Turkish.

*** Duration: July 15, 2020 - July 15, 2022**
RAs supported by the project: Pouya Ghahramanian, Sepehr Bakhshi
(Budget 263,219 TL ($38,337 as of July 15, 2020: official project beginning date)