Supervised and semi-supervised classification of concept drifting multi-label data data streams using novel neural network approaches *

TÜBİTAK Project No. 120E103

Fazlı Can (PI), Seyit Koçberber (I)

December 15, 2022

Topic. Data-stream classification is a major active research area that has gained importance with the emergence of continuous stream producing resources such as web, IoT and social media. In these domains data items can be related to several classes. In data stream processing, the classification should be done instantly by using limited resources in the arrival order of data items and by dynamically changing the learning model. Furthermore; the concept drift, change in the relationships between the data items and the classes assigned to these items, necessitates the adaptation of the learning model to these new conditions. The classification of evolving multi-label data streams using supervised and semi-supervised new neural network approaches is our research topic in this project.

Contributions. In our project the BELS neural network that we derived BLS will be adapted to the multi label classification problem (1st contribution: ML-BELS - we provide the first BLS-based multi-label classifier for data-streams). The neural-network structure of the first step will be transformed into a semi-supervised neural-network structure (2nd contribution: S2ML-BELS - we provide the first self-training neural network structure for multi-label classification in data streams). A new method that works in a supervised or unsupervised manner for detecting the concept drifts will be developed (3rd contribution: CCDD). The three parts of our study will be evaluated by statistical methods in a wide range of complementary experiments.

Method. BELS (Broad Ensemble Learning System) neural network has been defined by our research group with theoretical derivations is the first and only approach that adapts the BLS (Broad Learning System) neural network structure developed for the classification of traditional datasets to data streams. Different from traditional multi-layer approaches BLS involves one broad layer and by this way it provides effectiveness and efficiency. In multi-label BELS (ML-BELS) multiple components of BELS will be trained for different labels and the final decision will be determined by the weighted average of results. The semi-supervised multi-label BELS structure, S2ML-BELS, learning will be

provided using a limited number of labeled data. In both approaches, the worsening effects of concept drift will be eliminated by a pool that stores the previously acquired learning models. Our concept drift detection method proposed in the project, CCDD (Clustering-based Concept Drift Detection), will detect the shift by tracking the change in the coexistence of labels assigned to data items.

Project Management. The project involves five work packages (WP 1-5) as follows: Preparation and generation of experimental data sets (WP 1); Development of a broad neural network structure that performs supervised (WP 2: ML-BELS) and semi-supervised (WP 3: S2ML-BELS) classification in multi-label data-flow environments; Design and construction of a new algorithm that detects concept drift in multi-label dynamic classification environments (WP 4: CCDC); and The evaluation of the methods we have developed with comprehensive experiments and statistical tests, together with the baselines, and making the approaches we have developed more effective and efficient based on the observations (WP 5)

Impact. The ML-BELS ve S2ML-BELS neural network structures that we define have the potential of being a new baseline algorithms in the literature, just like our previous data stream research results. The approaches we develop can be adapted to classification problems that involve a very large number of labels. Our supervised and semi-supervised neural network structures can be used in a wide range of applications as they transform and complement each other. Multi-label experimental datasets containing concept drifts that we will produce for Turkish during the project would provide opportunities for researchers who are new to the topic.

*** Duration: December 15, 2022 - December 15, 2024**
RAs supported by the project: Pouya Ghahramanian, Sepehr Bakhshi
(Budget 486,274 TL (as of December 15, 2022: official project beginning date)