Online Classification of Multi-Source Parallel Data Streams In Big Data Environments*

Büyük Veri Ortamlarında Çok-Kaynaklı Veri-Akışlarının Çevrimiçi Sınıflandırılması

TÜBİTAK Project No. 117E870

Fazlı Can (PI), Seyit Koçberber (I)

June 1, 2018

Abstract

Data streams are essential component of big data domains that contain a countless number of data items in time order. In this project, we will examine the classification problem in the data stream environments in a new flexible framework that observes multiple sources of data streams and we believe that we will make significant contributions to the literature.

Along with advances in technology, and also due to the automation of several daily processes, an enormous increase is observed in the variety, velocity, and volume of data streams. Examples of data streams in the context of big data include social media posts, traffic sensor data, electronic transactions related to stock exchanges and banks, intelligence reports and observations, recommendations to customers in shopping sites , IOT communication data, and medical measurements.

Relevant data stream applications range from increasing user satisfaction in social media, to making decisions in vital situations. The evolving nature of data streams, and the necessity of making correct predictions with limited time and memory, based on a select number of recent data items, , make data stream mining a challenging and interesting field of research. Applications based on research findings improve quality of life, and are an important resource of income for the industry.

In the literature, it is traditionally assumed that there is only one data stream. On the other hand, in many everyday applications, there are multiple parallel data streams of data originating from multiple sources regarding a particular classification problem. For example, with respect to credit card usage, there are data streams that naturally reflect different spending habits from various settlements. The merging of these data streams lead to the disappearance of individual aspects, and lowers the accuracy of predictions. However, there is a limited number of studies on multi-source data streams. This can be due

to the fact that only recently there are some software development platforms, like SAMOA, for distributed data stream processing.

In data mining, it is known that the aggregated prediction obtained with an ensemble of classifiers is better   than the best individual component. In our recent work, we developed an algorithm called GOOWE, which aims to give the best estimate in a single-source data flow environment;it combines the predictions of component classifiers with a geometric approach, in a multi-dimensional space,. The theorem we proved shows that the number of class labels of the dataset is also the ideal number of component classifiers, with the premise that they generate independent scores. An ensemble with a larger number of members, however, only leads to resource waste, following many examples of the "diminishing returns" law. These findings will be employed in a novel two-stage framework by adapting GOOWE in parallel multi stream data environments. The new framework is adaptable to various real-life applications.

We can summarize the contributions of this study under four headings: In this work we 1. Propose a new data stream classification approach based on multiple data streams different from traditional single stream-based approaches. Design a new two-stage classification framework for multiple data stream environments; 2. Reflect the learning outcomes obtained by GOOWE for different data streams, related to the same classification problem to an active, higher-level ensemble algorithm proposed in this project; 3. Establish the framework and reflect the theory that we have developed to practice for multiple parallel data streams in the SAMOA platform; 4. Demonstrate the scalability of our proposed framework, and its suitability for large data mining and prove its success with experiments and statistical tests.

The flexibility of the framework we recommend makes it adaptable to different classification problems, and provides new research opportunities. The algorithm we propose to use in decision making by the higher-level ensemble is novel; it is an active algorithm that can adapt itself to incoming data streams and can also be easily adapted to multi-source data streams flowing with different speeds. GOOWE, which adapts itself to the observed stream in a reactive way, without parameters, will be used for the first time by adapting it to multi-stream environments. Scalability in data stream classifier studies is an important step in transferring research results into real life. The framework we will develop using the SAMOA platform will be an example for other researchers.

*** Duration: June 1, 2018 – June 1, 2020**
RAs supported by the project: Aykut Güven (MS Student), Ömer Gözüaçık (MS Student)
Budget 206,250 TL ($44,832 as of June 1, 2018: official project beginning date)