Lifelong Machine Learning on Data Stream

The interest in developing machine learning (ML) algorithms for streaming data research has increased over the years. Complex problems provide data with new characteristics in comparison to the previous decade. Nowadays, data sources generate high dimensional information in large volumes and velocity. The generated data are becoming increasingly ubiquitous. To extract useful information is a real challenge. The provided data does not have a homogeneous structure, and it can also be redundant, noisy, and incomplete. These characteristics have motivated the development of several machine learning algorithms for data streams analysis. Streaming data research has mainly focused on developing accurate decision models with the ability to learn and forget concepts incrementally. The advances in streaming data research have been useful in areas such as clustering, temporal learning, anomaly detection, semi-supervised learning, novel class detection, and feature selection. ML research for data streams often appears on the big data (BD) analysis domain, related to tackling issues concerning to velocity and volume of the BD.

The main characteristic of the data stream classification is the possibility of the large amount of data appearing sequentially, creating endless data stream over which the observer does not influence the order of incoming instances. Moreover, a classifier has to be ready at all times to make a decision. During the continuous classifier update step, one has to take into consideration the limited memory and computational resources. Furthermore, we may be faced with non-stationary data streams, i.e., where the data distributions may change, forcing the classification model to adapt to upcoming changes. This phenomenon is called *concept drift*, and its nature can vary due to both the character and the rapidity. It forces the implementation of mechanisms enabling adapting to the current class imbalance or concept drift detectors that providing a drift occurs enforces the model to be rebuilt. Another important issue is the availability of the class labels. Many techniques assume that labels are always at disposal or with a slight delay. Unfortunately, data labeling is connected with a sizable cost; therefore, it's naive to presume that one can have full knowledge of the class labels. As a result methods using (*semi-supervised learning*) and (*active learning*) are gaining significant popularity.

Lifelong ML systems can overcome limitations of statistical learning algorithms that need a large number of training examples and are suitable for isolated single-task learning. However, existing lifelong ML research is still in its infancy, and there are many open challenges. Key features that need to be developed within such systems to benefit from prior learned knowledge include feature modeling, knowledge retaining from past learning tasks, knowledge transfer to future learning tasks, previous knowledge update, user feedback. Also, the concept of "task" that appears in many formal definitions of lifelong ML models, seems to be hard to define in many real-life setups (it is often difficult to distinguish when a particular task finishes and subsequent starts). One of the main challenges is the *stability and plasticity* dilemma, where the learning systems have to trade-off between learning new information without forgetting the old one. It is visible in the catastrophic forgetting phenomenon, which is defined as a complete forgetting of previously learned information by a neural network exposed to the new information. Another open challenge refers to the evaluation of lifelong ML methods. The classification of data streams usually presents the case of large flows of data, among which specific events appear from time to time. Statistical knowledge about them can help the design of classifiers for these types of tasks. The main problem here is the fact that the rarer the occurrence of the targeted situation, the more costly it is to extract information from data. Simulated data will exhibit the same problem. Therefore, we plan to incorporate streams with rare events in our empirical evaluations.

In order to provide research contributions, we structured this three-year project with the following specific objectives:

A. New development in the drift concept area.
B. Advanced streaming data research using recurrent systems.
C. Lifelong learning on streaming data.

This project involves researchers from Faculty of Electrical Engineering and Computer Science (FEI), VŠB-Technical University of Ostrava and the Department of Systems and Computer Networks, Wroclaw University of Science and Technology.