Creating a Flink Adapter on Apache SAMOA

Apache Scalable Advanced Massive Online Analysis (SAMOA) is a platform for mining data streams with the use of distributed streaming Machine Learning algorithms, which can run on top of different Data Stream Processing Engines (DSPEs).

As depicted in Figure 20, Apache SAMOA offers the abstractions and APIs for developing new distributed ML algorithms to enrich the existing library of state-of-the-art algorithms [27, 28]. Moreover, SAMOA provides the possibility of integrating new DSPEs, allowing in that way the ML programmers to implement an algorithm once and run it in different DSPEs [28].

An adapter for integrating Apache Flink into Apache SAMOA was implemented in scope of this master thesis, with the main parts of its implementation being addressed in this section. With the use of our adapter, ML algorithms can be executed on top of Apache Flink. The implemented adapter will be used for the evaluation of the ML pipelines and HT algorithm variations.

5.1 Apache SAMOA Abstractions

Apache SAMOA offers a number of abstractions which allow users to implement any distributed streaming ML algorithms in a platform independent way. The most important abstractions of Apache SAMOA are presented below [27, 28].