

GE461: Introduction to Data Science

Assignment for Data Stream Mining

May 12, 2023

Due date: June 2nd, 2023; 11:59 pm

Notes: In this assignment, you will build an evolving data stream classification model. You will use scikit-multiflow, an open-source Python library that provides tools for data stream mining and classification. You can find more information about this library on the official webpage (<https://scikit-multiflow.github.io/>). Your task is to train data stream classification methods on the given data streams. You will use test-then-train –i.e., prequential evaluation method to compare performance of the classification models.

Teaching Assistant: Sepehr Bakhshi, sepehr.bakhshi@bilkent.edu.tr

TA Office Hour Administration: Please send your question to Sepehr. He can arrange a meeting upon request.

A. What to Submit

Your submission has two components.

1. **Code.** It must contain proper comments. You must also include your name at the top as a signature that confirms that you are the programmer. Please remember that MOSS is in our plans for plagiarism check.
2. **Report.** Your report must be in pdf form and must cover all "Work to be Done" sections of the assignment (except part C.1 and C.2). Explicitly write the steps that you have taken in each section. Please also write a brief introduction that includes the overall structure of the report (You can use introduction part of the scientific papers as an example to write a similar one for this homework). For each section of your work explain the purpose and what has been done and achieved in that section. Provide a comparison of results that contains tables and plots as appropriate. Make sure that you follow the principles of scientific writing. Use simple past or simple present tense in your report. If you plan to propose future work then in that case you may use future tense.

Your report must have proper title like a scientific paper, reflecting its true content. It must have your name and address etc. If you like, for experience and fun, you may use the ACM conference paper format¹. You must use latex or Microsoft Word or their equivalent.

Optional: As an optional part, at the beginning of your report, you may have a related works section that covers data stream mining briefly with proper references.

Optional: Another optional part is comparison of the effectiveness of the methods using statistical tests. The design and administration of these tests should be decided by you by looking at the available papers in literature.

See Justin Zobel's book *Writing for Computer Science* for further hints on the style of CS related scientific paper writing.

B. Submitting Your Work

You will submit your work by uploading it to Moodle in a zipped file. Its name must be streamMiningYourFirstNameYourLastName. For a student with the name "Ali Can Ok" it is streamMiningAliOk.

C. Work to be Done

1. Concepts:

- ***Data Stream:*** refers to an environment where data arrives continuously over time. This means we cannot assume that we have access to all the data at the beginning. Instead, we need to update our model incrementally as new data arrives. This is in contrast to traditional batch learning, where we can access all the data simultaneously and train the model on the entire dataset.
- ***Concept drift:*** refers to a change in the underlying distribution of the data. For example, in the case of a spam email detection model, the characteristics of spam emails may change over time, which means that the system needs to adapt and update itself to identify the new characteristics of spam emails correctly.

¹ https://www.acm.org/binaries/content/assets/publications/taps/acm_layout_submission_template.pdf

Data stream classification models need to adopt a concept drift detection and handling method to address concept drift. We refer to a data stream with concept drift as evolving data stream.

- **Prequential Evaluation:** is a commonly used evaluation approach for data stream classification tasks. In the prequential evaluation method, the model is tested on each incoming instance before it is used to update the model. It is also known as interleaved test-then-train evaluation method.

2. Requirements:

You will need the following libraries in Python to complete this challenge.

- numpy: a Python library for numerical computing
- scikit-learn: a library for machine learning in Python
- scikit-multiflow: a library for data stream mining and classification

3. Datasets as Data streams

You will use two synthetic and two real datasets as data streams to compare performance of the classification models. As synthetic data streams, use AGRAWALGenerator and SEAGenerator classes from scikit-multiflow to generate 100,000 data instances for each. For future access, write the generated data instances into files named AGRAWALGenerator and SEADataset.

To apply concept drift on these datasets, please read the documentation of [SEAGenerator documentation](#)

As real datasets, you will experiment with the Spam and Electricity datasets. You can obtain these datasets from <https://github.com/ogozuacik/concept-drift-datasets-scikit-multiflow>.

4. Classification Task with Concept Drift Handling

4.1. Implement an instance of the following classification models. You can use scikit-multiflow to this aim.

- Adaptive Random Forest (ARF) [4]
- Streaming Agnostic Model with k-Nearest Neighbors (SAM-kNN) [3]
- Streaming Random Patches (SRP) [1]
- Dynamic Weighted Majority (DWM) [2]

4.2. Build your own ensemble from scratch. Use HoeffdingTreeClassifier classifier as your base learner.

Find a solution for handling drift in your ensemble and implement it.

Hint: You can use drift detection mechanisms also included in the scikit-multiflow.

Note: The ensemble approach should be implemented from scratch. You are only allowed to import the HoeffdingTreeClassifier, your preferred drift detector, and other basic libraries (numpy, pandas, etc.)

5. Results and Discussion

Construct an instance of the classification models. For each dataset, use Interleaved Test-Then-Train approach to train and evaluate performance of these classifiers. Use prediction accuracy as evaluation metric. Report the following results for the classification models on each dataset:

- Overall accuracy: Overall prediction accuracy of the models.
- Prequential accuracy plot: Prequential accuracy is defined as the prediction accuracy of a model over the w most recent data instances. Use 20 sliding windows of size (dataset size/20) to calculate prequential accuracy values. Plot the obtained accuracy values over time for each dataset.

5.1. How does your ensemble model perform compared to the state-of-the-art approaches in 4.1? What could be possible improvements for a more robust ensemble.

5.2. Discuss your findings on the accuracy plots. What is inferred from the drops in the prequential accuracy plot?

5.3. Please also include a paragraph that summarizes your findings in this assignment. What did you learn from this assignment?

In your comparisons use plots and tables when appropriate. Number all plots and tables and provide proper subtitles for them. Make sure that you refer to each of them in the text of your report. Help your reader by providing a simple and easy to follow presentation.