University of Ottawa

School of Electrical Engineering and Computer Science

CSI5155 - Fall 2022

Assignment 4: Online Learning

TOTAL MARKS 90

Instruction:

1. This is an individual assignment. Submit your assignment using BrightSpace, before the due date.
2. For the implementation, you should either upload your code on BrightSpace or provide a link to a GitHub repository. Note that, if you choose to use GitHub, the date and time of last change to your repository should be **before** the assignment deadline.
3. Use Scikit-Multiflow [2, 3] to complete the assignment.

The aim of this assignment is to obtain a first exposure to online learning against an evolving data stream, within the Scikit-Multiflow environment [2, 3]. Specifically, we are studying the impact of concept drift on algorithm behaviour, whilst conducting prequential evaluation.

You are asked to use three of the Insects data streams [1] in this assignment. The data may be downloaded at https://sites.google.com/view/uspdsrepository (the password is **DMKD2018**).

This data was obtained from a laser sensor built with low-cost components to remotely capture information about flying insects, in order to aid in intelligent insect trap design [1]. Specifically, we will **only** use the Insects-Abrupt-Balanced, Insects-Incremental-Balanced, Insects-Gradual-Balanced streams in this assignment. Sections 5 to 7 of [1] contain details about the data and the experimentation relevant for this assignment.

1. In Section 7.1 of the reference paper [1], the authors first consider the no-change and majority class classifiers, with a moving window over a stream of 1000 instances. As a first step, you are asked to conduct these experiments against the three data streams listed above. Following [1, 4], use prequential accuracy over a sliding window of 1000 to report your results.

[5 marks]

2. Next use the following algorithms to construct models against the three data streams: Hoeffding Trees, SAM-KNN, Hoeffding Adaptive Trees as well as two (2) ensemble-based methods of your choice. Again, you should report the prequential accuracies over a sliding window of 1000 instances. `` **[25 marks]**
3. Create figures, similar to figures 22 to 27 in [1], to show the prequential accuracies against the three streams, for the learners used in steps 1 and 2. [**5 marks**]
4. Next, combine the Hoeffding Tree learner with a drift detection method of your own choice, again using the same setting as the paper in terms of window size (1000). Report the prequential accuracies over a sliding window of 1000 instances. **[10 marks]**
5. Create figures, similar to figures 28 to 30 in [1], to show the prequential accuracies against the three streams, for step 4. [**5 marks**]
6. Create a table, similar to table 5 in [1], summarizing the prequential accuracies you achieved in steps 1, 2 and 4. [**5 marks**]
7. Discuss the results you obtained and the lessons you learned when analysing this data. **[20 marks]**
8. Contrast the results you obtained during this assignment with those of the reference paper [1]. Be sure to discuss any differences in methodologies, and results, and to highlight similarities. **[15 marks]**

References

[1] Souza, V.M.A., dos Reis, D.M., Maletzke, A.G. *et al.* Challenges in benchmarking stream learning algorithms with real-world data. Section 5, *Data Mining and Knowledge Discovery,* **34**, 1805–1858 (2020).
URL: https://link.springer.com/article/10.1007/s10618-020-00698-5

[2] Scikit-Multiflow, URL: https://scikit-multiflow.github.io/

[3] Scikit-Multiflow learning methods,
URL: https://scikit-multiflow.readthedocs.io/en/stable/api/api.html#learning-methods

[4] Bifet, A., Gavaldà, R., Holmes, G., and Pfahringer, B. Machine Learning with Data Streams with Practical Examples in MOA, 2018:
URL: https://moa.cms.waikato.ac.nz/book/