Supplementary Material of Online Ensemble Model Compression for Nonstationary Data Stream Learning

Rodrigo G. F. Soares, Leandro L. Minku

I. Tables with accuracy results

Figure 1 depicts the plots of G-Mean of each compared method across several numbers of base learners on each data stream separately using 10, 20, 30, 40 and 50 base learners.

Figure 2 depicts the plots of G-Mean of each compared method across several numbers of base learners on each real-world data stream separately using 10, 20, 30, 40 and 50 base learners.

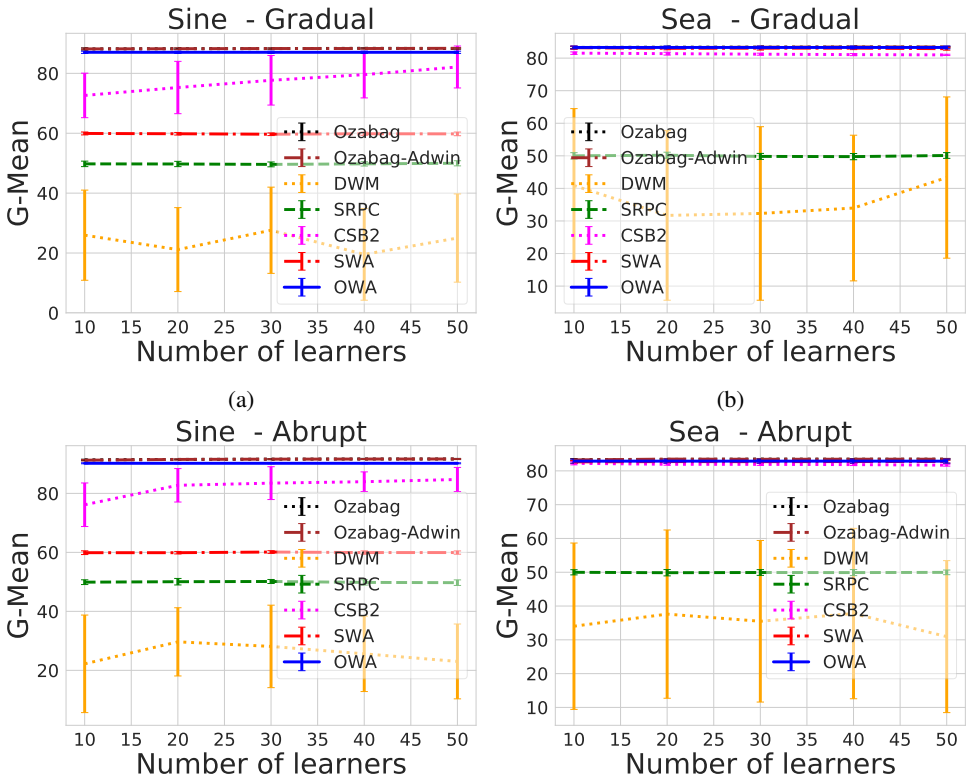

Fig. 1: Plots of G-Mean measured over 30 replicates of each method using different number of base learners on artificial data streams.

II. Experiments with visual data

Our approach is not designed for specific problem of image classification, nevertheless we have also run experiments with four image datasets to evaluate its predictive performance on this particular problem.

Rodrigo G. F. Soares is with the Department of Statistics and Informatics of the Federal Rural University of Pernambuco, Recife, Brazil and with the School of Computer Science, The University of Birmingham, Birmingham B15 2TT, UK. rodrigo.gfsoares@ufrpe.br.

Leandro L. Minku is with the School of Computer Science, The University of Birmingham, Birmingham B15 2TT, UK. L.L.Minku@cs.bham.ac.uk.

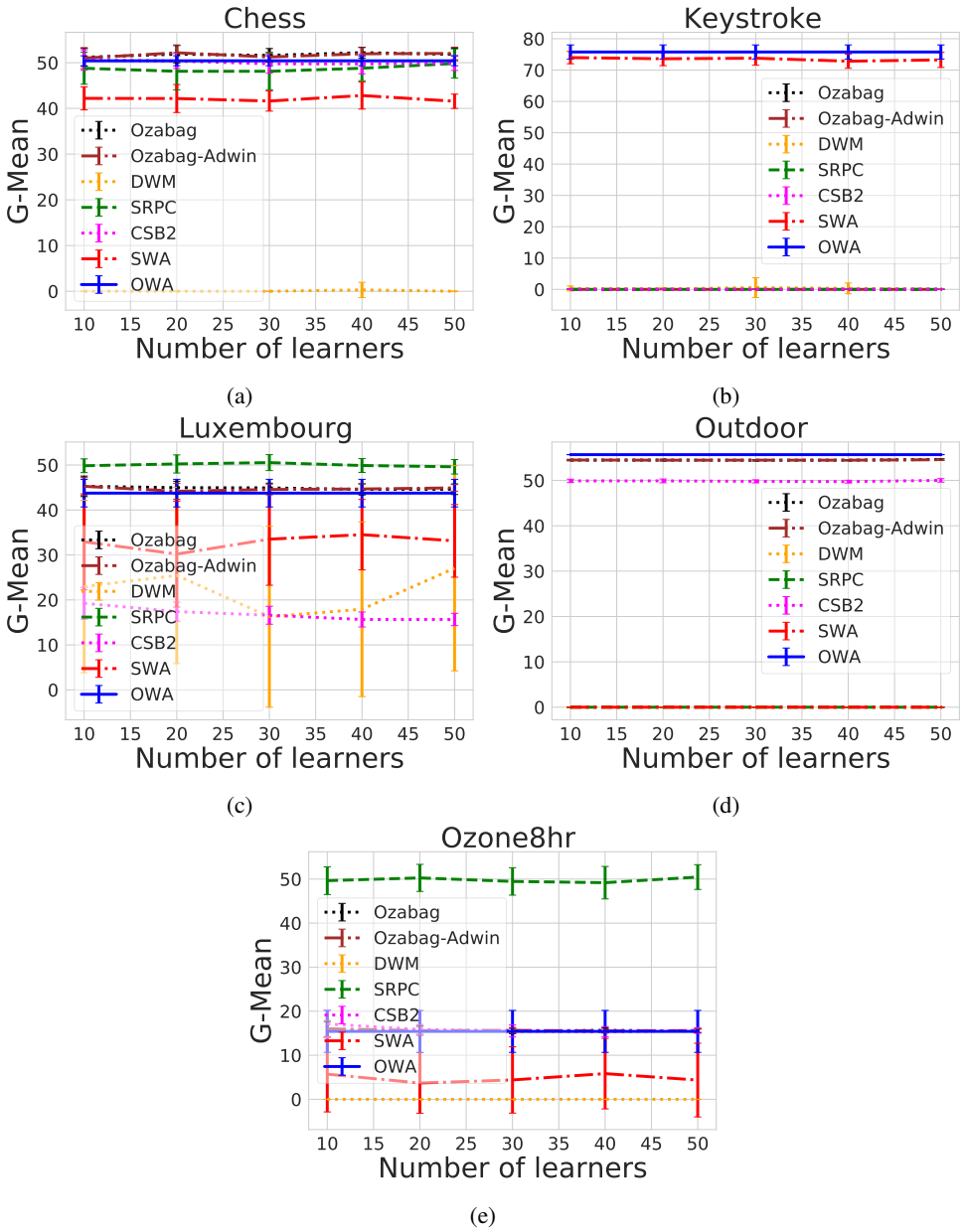

Fig. 2: Plots of G-Mean measured over 30 replicates of each method using different number of base learners on real-world data streams.

We used 4 data streams: Outdoor [2], [3], Rialto [3], CIFAR [1] and Rotated MNIST [4], where the instances in Outdoor and Rialto are naturally ordered forming a true data stream. CIFAR and Rotated MNIST are datasets typically used for offline learning, as they have images in randomized orders. However, they were used to simulate data streams by presenting such images sequentially to the machine learning approaches. The summary of these streams is as follows:

- **Outdoor** [2], [3] contains visual data that was captured from images recorded by a smartphone camera in a garden environment. There are 40 different classes of objects to be classified: balls, shoes, pliers, cans, among others. Each object has 100 images produced under varying lighting conditions (sunny and cloudy) and from different distances and positions. There are 4,000 images recorded and arranged in temporal order. Each example is represented by a normalized 21-dimensional RG-Chromaticity histogram.
- **Rialto** [3] contains color images extracted from time-lapse videos recorded by a webcam in a fixed position. The recordings cover 20 consecutive days from May to June 2016, capturing various colorful buildings next to the famous Rialto bridge in Venice. We employed the buildings number 0 and 4 were considered as the classes for our classification problem. It has 16,450 instances and 27 features.
- **CIFAR** contains resized (16x16 pixels) gray-scale images from the original CIFAR10 dataset [1]. We selected instances

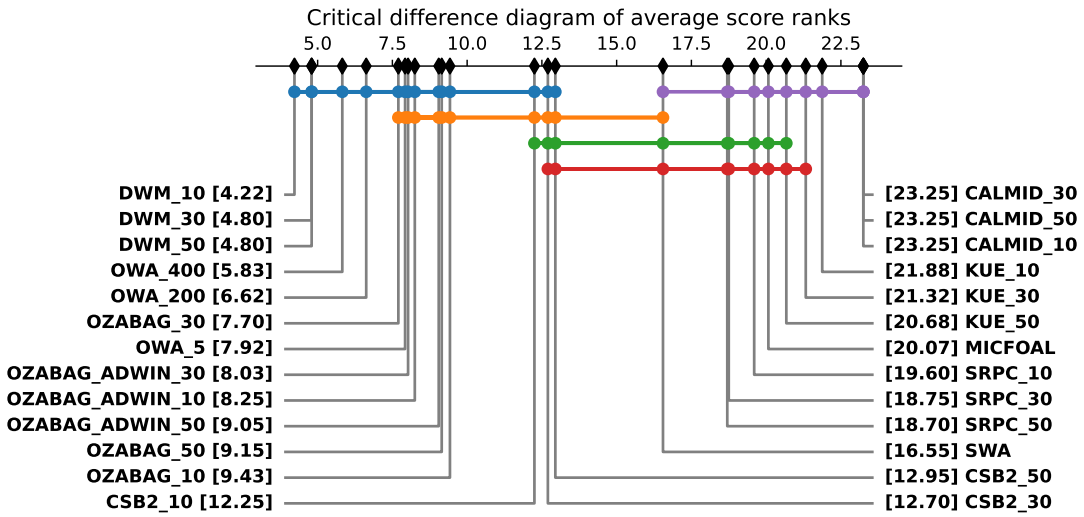

Fig. 3: G-Mean Nemenyi post-hoc statistical test across artificial data streams. For ensemble methods, the numbers of base learners are 10, 30, 50. For OWA, $C \in \{5, 200, 400\}$. Values in brackets are mean ranks, where smaller ranks represent higher G-Means. Groups of algorithms that are not significantly different are connected with a bold solid line.

from classes "automobile" and "dog". This stream has 12,000 instances and 256 features.

- **Rotated MNIST** [4] consists of rotated resized (16x16 pixels) gray-scale images of handwritten "0" and "1" digits. It has 12,670 instances and 256 features.

We analyze the results for G-Mean calculated prequentialy. We used the Friedman statistical test followed by the Nemenyi post-hoc test with significance level of 0.05 to evaluate statistical differences in G-Mean. Figure 3 depicts the critical difference diagram for the Nemenyi post-hoc tests to identify which methods perform differently from each other. We compared ensembles with different numbers of base learners. We use a $C = 5$ for OWA. Based on these tests, none of the compared methods produced significantly superior predictive performance to OWA across artificial data streams. In particular, OWA was significantly better than SWA and no significant difference has been found compared to MICFOAL, KUE, CALMID and SRPC.

In Figure 4, we show the heatmap for all methods on visual data. Ensemble methods are shown with different numbers of base learners (10, 20, 30, 40, 50). OWA is shown with lengths of the cycle $C \in \{5, 50, 100, 200, 400\}$. For every data stream, OWA produced similar or better G-Mean to the other methods with different numbers of base learners.

The Figure 5 shows the G-Mean over time of the best run of each algorithm on streams with visual data. OWA is able to deliver superior performance throughout the Outdoor and Rialto streams. For CIFAR10 and RMNIST, several methods have comparable generalization, except CALMID, KUE, MICFOAL and SWA. This demonstrates that OWA delivered G-mean similar to the best ensemble methods over time. This result supports OWA as an effective alternative to ensemble methods without the computational overhead of training multiple learners. It is important to highlight that the compared ensemble methods have sophisticated learning mechanisms (for example, DWM creates and removes base learners as necessary), whereas OWA achieves similar or better performance with a simple weight averaging through time that resets its weights to $\mathbf{w}$ when a concept drift is detected.

References

[1] Alex Krizhevsky. Learning multiple layers of features from tiny images. Technical report, University of Toronto, 2009.

[2] Viktor Losing, Barbara Hammer, and Heiko Wersing. Interactive online learning for obstacle classification on a mobile robot. In *2015 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8, Killarney, Ireland, July 2015. IEEE.

[3] Vinicius M. A. Souza, Denis M. dos Reis, André G. Maletzke, and Gustavo E. A. P. A. Batista. Challenges in benchmarking stream learning algorithms with real-world data. *Data Mining and Knowledge Discovery*, 34(6):1805–1858, 2020.

[4] Rupesh K Srivastava, Jonathan Masci, Sohrob Kazerounian, Faustino Gomez, and Jürgen Schmidhuber. Compete to compute. In C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 26. Curran Associates, Inc., 2013.

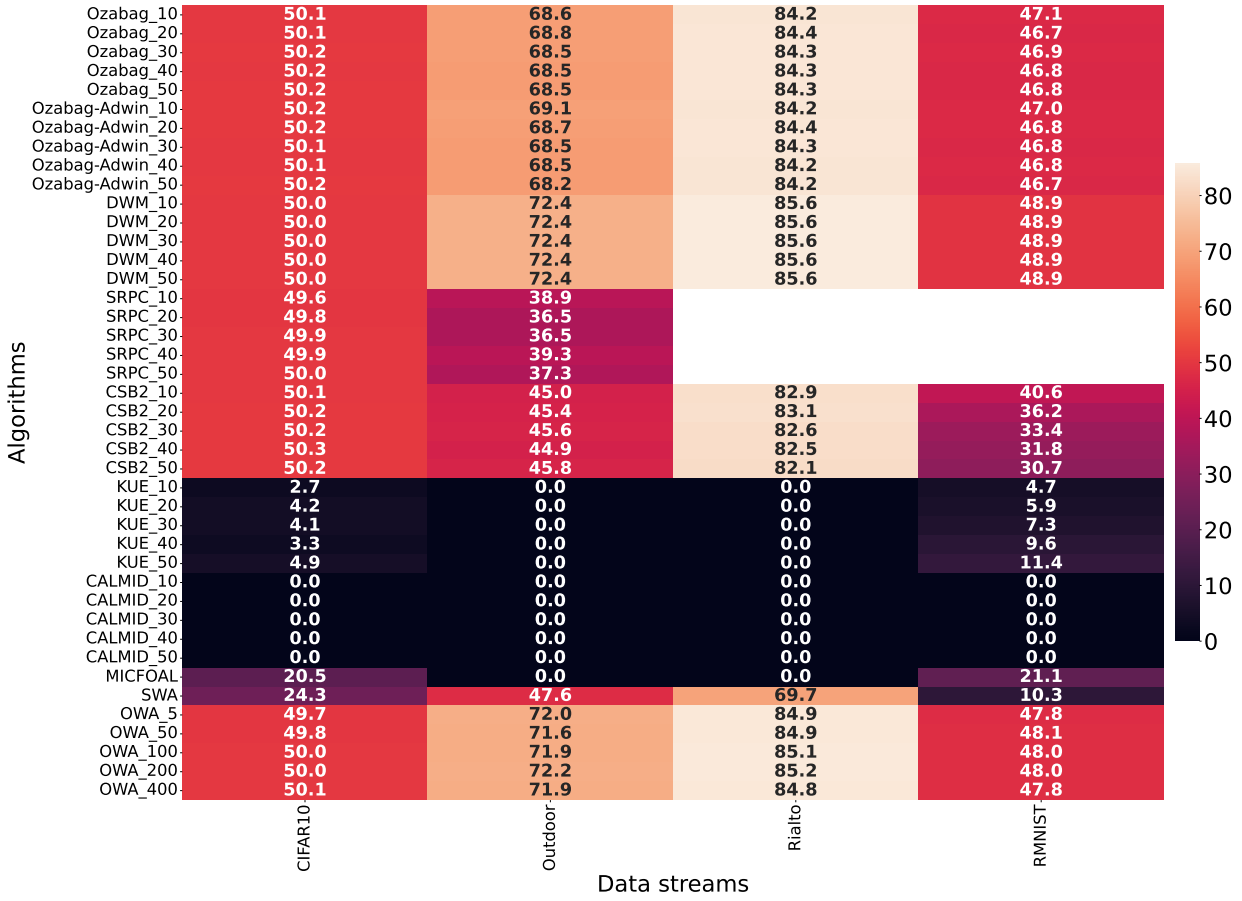

Fig. 4: Heatmap of G-Mean for all methods on visual data. Ensemble classifiers are shown with different numbers of base learners. We show OWA with different lengths of cycle.

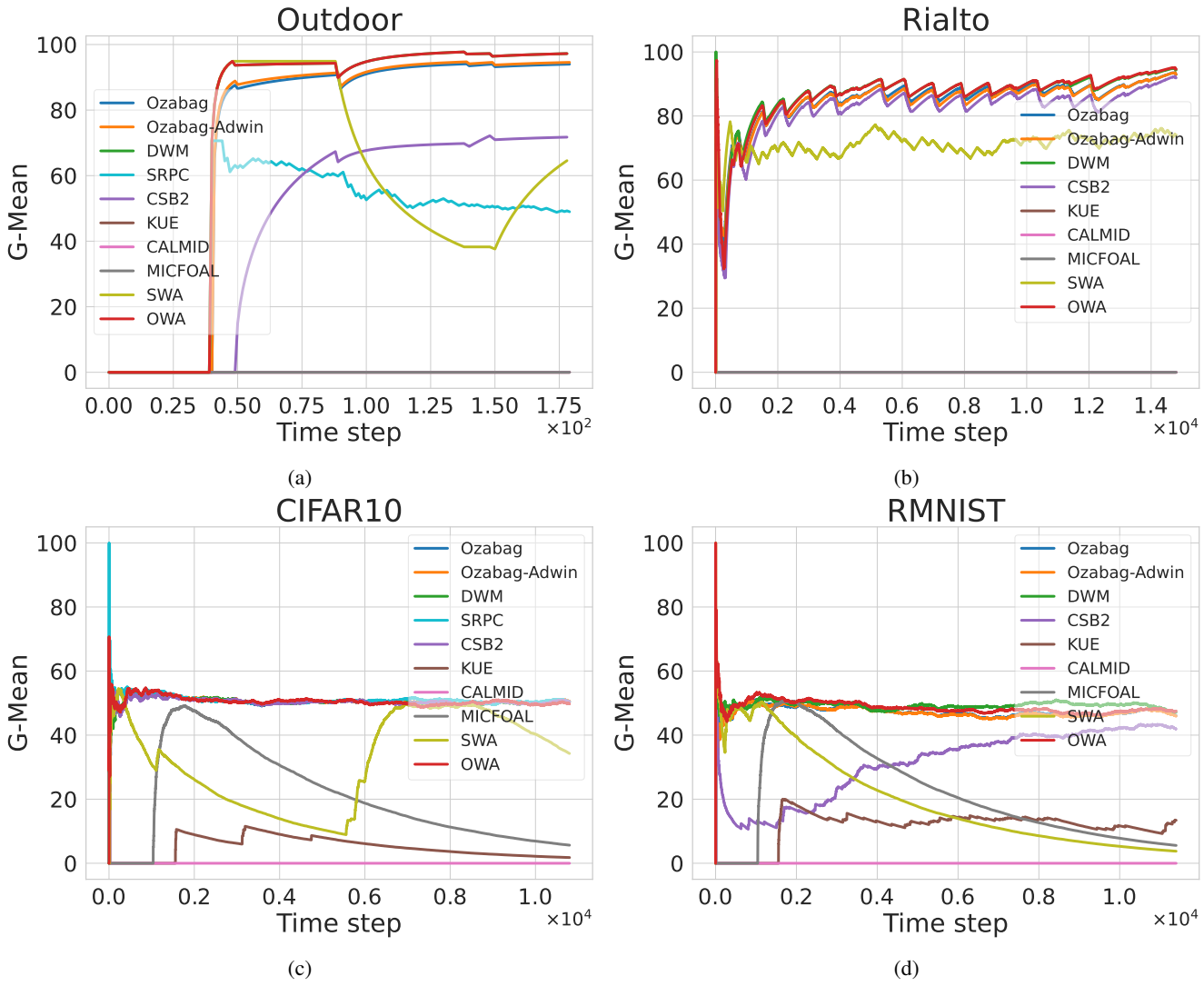

Fig. 5: Plots of G-Mean produced by the best run of each of the methods for visual data.