Supplementary of "A Semi-Supervised Active Learning Neural Network for Data Streams with Concept Drift"

Botao Jiao, *Student Member, IEEE,* Heitor Murilo Gomes, Bing Xue, *Senior Member, IEEE*, Yinan Guo*, *Member, IEEE*, Mengjie Zhang, *Fellow, IEEE*

S1. Benchmark Datasets

The characteristics of data generators and real-world data streams are illustrated as follows:

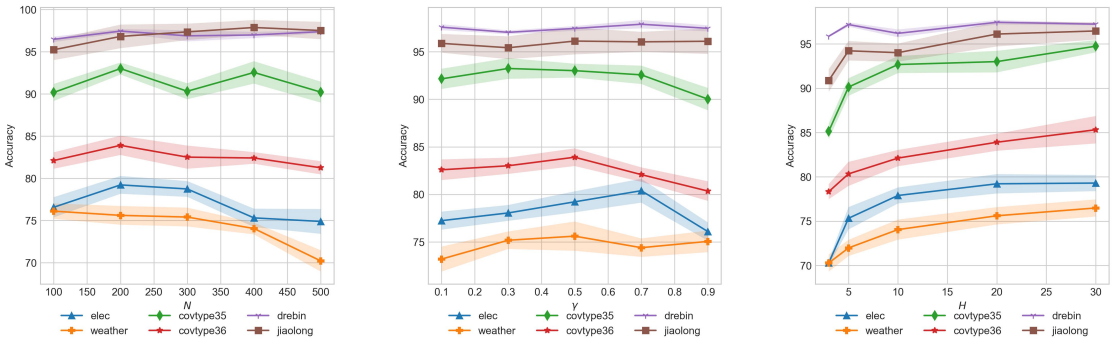
1) AGRAWAL generator [1]: It produces a stream with six numeric features and three categorical ones. Since non-tree classifiers cannot handle categorical features, each categorical feature with K values is converted into K binary features. In the experiments, four functions are employed to simulate sudden drift and form two data streams. Each data stream contains 5% noise.
2) Sine generator [2]: It creates instances with two relevant attributes that distribute in $[0, 1]$. The label of each instance is determined by a sine curve. As a concept drifts, the labels of two classes are reversed. Two data streams are generated in the similar way to AGRAWAL generator. This generator does not contain a function to generate noise.
3) RBF drift generator [3]: t randomly creates centroids and assigns the corresponding standard deviations, labels and weights. A new instance is created by randomly choosing a centroid and offsetting the attributes in a random direction based on a radial basis function. The chosen centroid also determines the label of a new instance. Concept drift is introduced by changing the centroids and the corresponding standard deviations. Two data streams with gradual drift are generated by this generator.
4) RandomTree generator [1]: A stream is generated based on a tree that is constructed by splitting features randomly and assigning labels at their leaves. After setting a random value satisfying uniform distribution to each attribute, a new instance is created. New concepts are generated by creating new trees. This generator produces data stream and concept drift occurs in it by a similar way to RBF generator.
5) Hyperplane generator [4]: The d-dimensional instances generated by Hyperplane satisfying $\sum_{i=1}^{d} w_i f_i = w_0$ where f_i represents ith feature of an instance and w_i is its weight. The generator change weights to produce the orientation and position of a hyperplane, forming incremental drift. In the generated data streams, the magnitude of each change is set to 0.001, and the weight is updated by $w_i = w_i + 0.001 * \sigma$, where $\sigma \in \{1, -1\}$ and has a 10% probability of switching. Five percent noise is introduced.
6) Electricity[1] [5]: The Electricity dataset contains 45,312 instances, collected every 30 min from the Australian New South Wales Electricity Market between 7 May, 1996 and 5 December, 1998. A binary classification task is to predict a rise (UP) or a fall (DOWN) in the electricity price. The data is subject to concept drift caused by time-varying consumption habits, unexpected events and seasonality.
7) Weather[2] [1]: The Nebraska Weather prediction dataset is compiled by U.S. National Oceanic and Atmospheric Administration from 1949-1999, providing a wide scope of weather trends. It contains eight features and 18159 instances with 31% positive (rain) classes and 69% negative (no rain) ones.
8) Covtype[3] [6]: It contains 581,012 instances for predicting 7 classes of forest cover given by US Geological Survey (USGS) and US Forest Service (USFS). Each instance contains 54 geographic features, such as Elevation and Slope.
9) Drebin [7]: This dataset is designed to identify Android malware and comprises ten textual attributes, such as API calls, permissions, and URLs. It consists of 123,453 benign and 5,560 malicious Android applications. In our analysis, we utilize the digital features extracted by Ceschin et al. [7] using the Term Frequency-Inverse

B. Jiao and Y. Guo are with School of Information and Control Engineering, China University of Mining and Technology, Xuzhou 221116, China. Y. Guo is also with Key Laboratory of System Control and Information Processing, Ministry of Education, Shanghai, 200240, and School of Electromechanical and Information Engineering, China University of Mining and Technology (Beijing), Beijing 100083.

B. Jiao, H. Gomes, B. Xue and M. Zhang are with the School of Engineering and Computer Science, Victoria University of Wellington, PO Box 600, Wellington 6140, New Zealand

This work is supported by National Natural Science Foundation of China under Grant 61973305, 61573361, 52121003, the Foundation of Key Laboratory of System Control and Information Processing, Ministry of Education, P.R. China under Grant Scip202203, National Key R&D Program of China under Grant 2022YFB4703701, the Marsden Fund of New Zealand Government under Contracts VUW1913, VUW1914, VUW2213, MBIE Data Science SSIF Fund under the contract RTVU1914, Huayin Medical under grant E3791/4165, and MBIE Endeavor Research Programme under contracts C11X2001, UOCX2104. The work of Botao Jiao was supported by the China Scholarship Council under Grant 202206420117

[1] https://moa.cms.waikato.ac.nz/datasets/
[2] https://users.rowan.edu/ polikar/nse.html
[3] https://archive.ics.uci.edu/ml/datasets/Covertype.



(a) N (b) γ (c) H

Fig. S1. Accuracy of SSALD with different values of parameters on real-world data stream

Document Frequency (TF-IDF).

10) JiaoLong [8]: The JiaoLong deep-sea manned submersible dataset is provided by the Qingdao National Deep Sea Center, China. It originates from the 151st exploration mission of the JiaoLong manned submersible on March 19, 2017. All data monitored during approximately 9 hours is presented in the form of multivariate time series. This dataset contains 14 features, among which 8 ones come from the internal life support system, while the rest are from the external monitoring sensors and the hydraulic system.

REFERENCES

[1] A. Liu, J. Lu, and G. Zhang, "Diverse instance-weighting ensemble based on region drift disagreement for concept drift adaptation," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 32, no. 1, pp. 293–307, 2021.

[2] J. Gama, P. Medas, G. Castillo, and P. Rodrigues, "Learning with drift detection," in *17th Brazilian Symp. Artificial Intelligence*, vol. 8, 09 2004, pp. 286–295.

[3] Y. Song, J. Lu, A. Liu, H. Lu, and G. Zhang, "A segment-based drift adaptation method for data streams," *IEEE Transactions on Neural Networks and Learning Systems*, pp. 1–14, 2021.

[4] A. Bifet, G. Holmes, R. Kirkby, and B. Pfahringer, "Moa: massive online analysis," *Journal of Machine Learning Research*, vol. 11, p. 1601–1604, 05 2010.

[5] J. Z. Kolter and M. A. Maloof, "Dynamic weighted majority: An ensemble method for drifting concepts," *Journal of Machine Learning Research*, vol. 8, no. 12, pp. 2755–2790, 2007.

[6] Y. Sun, K. Tang, L. L. Minku, S. Wang, and X. Yao, "Online ensemble learning of data streams with gradually evolved classes," *IEEE Transactions on Knowledge and Data Engineering*, vol. 28, no. 6, pp. 1532–1545, 2016.

[7] F. Ceschin, M. Botacin, H. M. Gomes, F. PinagÃ©, L. S. Oliveira, and A. GrÃ©gio, "Fast & furious: On the modelling of malware detection as an evolving data stream," *Expert Systems with Applications*, vol. 212, p. 118590, 2023.

[8] S. Hu, Z. Liu, M. Li, and X. He, "Cadm + : Confusion-based learning framework with drift detection and adaptation for real-time safety assessment," *IEEE Transactions on Neural Networks and Learning Systems*, pp. 1–14, 2024.