Review of: "A Novel Framework for Concept Drift Detection using Autoencoders for Classification Problems in Data Streams"

Hoang Anh Ngo[1]

1 Ecole Nationale Supérieure des Telecommunications de Paris

Potential competing interests: No potential competing interests to declare.

Comments on the overall content and format of the paper

The overall format of the paper should be improved, since various symbols and equations are extremely hard to read or understand, particularly Bayesian-related equations.

The authors have done a great and in-depth survey on supervised, unsupervised and semi-supervised concept drift detection models. This can potentially be improved to become a separate survey paper.

The authors should also consider making the code used in the experiments publicly available and provide a detailed guidance on how the results can be replicated. The repository https://github.com/Usman07442/AE-DDM-Experiments only contains information on the datasets used, not the actual implementation of the proposed framework.

Within their survey on concept drift detection models, there are discrepancies in the citation format. Some work are only mentioned by name, some are mentioned by the tuple (Authors, Year) and some are mentioned under the form of (Model's name, Year). This should be modified to avoid any ambiguity.

There are also various spelling, grammatical or lexical errors appearing here and there within the article. The authors should conduct proper checks for all of these errors before putting out another version of this article.

One major problem with this paper is that it is too long, and the main reason for that is the authors are doing the two enormous work in the same paper: a literature survey and proposal of a new algorithm. With that being said, the reviewer suggests that this paper should be divided into two smaller papers, one being a survey paper and one proposes the algorithm with more rigorous experiments.

Specific comments

The figure 1 that the authors opted to use to illustrate "real drift" and "virtual drift" does not fully represent the difference between the two. The previous paragraph should also be divided into bullet points with separated definition for each kind of drift to enhance readability.

Figure 2 also does not illustrate exactly different kinds of concept drift. Recurring concepts should have the recurrent factor, while the figure representing noise look very much the same compared to sudden drift.

At the end of page 4, when the authors say that “Supervised drift methods like DDM, RDDM, LFR and others […], these labels are not available immediately after the prediction most of the time and it is not possible to detect drift in real time using techniques”, this is not necessarily true. For example, with River, an online machine learning package written in Python, the algorithms are designed to provide labels immediately once the predict_one command is called, which allows such models to be effective in real time.

“Motived by …” should be changed to “Motivated by …”.

“The decoder part is the exact replica of the encoder part” makes sense in terms of idea, but is incorrect in terms of phraseology. The encoder part is the “mirror” of the decoder part, in which it decodes the smaller space into data that is exactly or as near to the original data as possible.

The function g(h) cannot be read properly.

The reviewer disagrees with what the authors mentioned in page 6 that “Another limitation of this work is that the same threshold is used for different datasets and we have experimentally observed that each dataset as well as each class data has its own reconstruction loss pattern and threshold.” Regarding online machine learning problems, we do not always know previously what is the distribution and characteristics of data streams look like; as such, having algorithms that work with various datasets under the same threshold would be the optimal path to follow.

Figure 7, AE-DDM framework within the proposed methodology are somewhat redundant. It only shows three separate part of the component without showing any relations or connections among the three, or between any two components. This figure should either be removed or improved.

In Table 5, why did the author opt to use this configuration? Is it proven to be optimal compared to all other configurations? Why should the bottleneck dimension be equal to one-third of the input dimension? Does MAE make any difference compared to MSE? What if with certain data streams, the ReLU activation function would be more optimal than the Sigmoid activation function?

Moreover, the reviewer think that instead of fixing one configuration/structure for the autoencoder, users should be allowed to have different choices based on their preference, experience and the specific data that they are dealing with.

Mean + 3*(standard deviation) should be written carefully, for example in math mode with the formula Mean + 3*SD to avoid any confusion.

The authors should also illustrate at which scale the reconstruction loss is compared to the original magnitude of the data points. A reconstruction error of 0.11 on data points with small and large magnitude are totally different scenarios.

How are the batch threshold and instance threshold values calculated for positive AE and negative AE in Stagger

dataset? This should be showcased explicitly within every experiment since the thresholds would have significant impact on the performance of the algorithm.

Minor remark: In figure 9, instead of writing “call algorithm XX”, the authors should instead state the purpose of the algorithm to make the flowchart more comprehensive.

In algorithm 6, why does the author fix he number of consecutive batches that exceed thresholds to 3? What would happen if this number is increased?

In the GAUSSIAN dataset, the mean and standard deviation of each feature should also be explicitly mentioned.

Minor remarks: in table 11, the average classifier accuracy should be expressed with 3 figures after the decimal point. Moreover, apart from accuracy, the authors should also consider reporting precision and recall, and any other metrics deemed necessary.

In table 14, is there any feasible explanation for the difference in performance with normal data and drifted data of the SVAM classifier?

Not only should the authors obtain results for their own algorithms, they should compare the performance also with other algorithms that are currently available. Moreover, apart from accuracy, the algorithm should take into account time complexity, which is an extremely important factor of any concept drift detection/online machine learning algorithm.

As mentioned previously, the future work including the extension to adapt with different autoencoder types should be integrated to this work for it to be more complete and robust.