Response to reviewers for manuscript APIN-D-19-01616

Henry Wilde, Vincent Knight, Jonathan Gillard

Included in this document are the responses of these authors to the comments made by the reviewers in regard to the manuscript submitted to Applied Intelligence as "Evolutionary Dataset Optimisation: learning algorithm quality through evolution".

Reviewer 1

> I think the subject and object of this paper is very ambiguous in the introduction even though authors well described the previous work, I hope the author should consider why this paper is necessary to the read.

The introduction has been revised to be more direct in its summary of the work, and its motivation has been expanded on. In particular, we have laboured the point even more that the proposed method belongs to a new paradigm in which no methods are currently published. We have expanded on the issues we raise with the current paradigm and provide further references to support our statements (see paragraph 4 of the introduction).

> The language needs to be revised by native speaker research.

The manuscript was written by three native English speakers but we have made sure that any lingering spelling mistakes or grammatical issues have been addressed. If there are any specific language errors that the reviewer can point out they will be addressed.

> make the ABSTRACT as a single paragraph and make sure do present your work in clearer.

In this draft of the paper, the abstract has been reduced to a single paragraph.

> You should improve the ABSTRACT.(rewrite the abstract to reflect the main idea and it's results, without any not suitable details) in the ABSTRACT alongside with the obtained results (the results you got it and what is the situation of your results in comparison with other published methods). Mentioned to the benchmarks which have been used in this paper.

We feel that the main concept of the method is summarised well in the abstract, including its motivation. As is discussed throughout this response, and in the article itself, classical results are not included in this work and no benchmarks are used hence those points are omitted from the abstract. However, we have revised the final sentence of the abstract to more accurately describe the case study provided at the

end of the article. In particular, we state that 'a number of known [favourable] properties' are identified by the proposed method for k-means and DBSCAN so that there is no ambiguity in the results displayed in the paper.

> in figure 1, there are question marks, you did not explain that. why?

The choice for question marks was to symbolise that a series of questions may be asked of the data. This has now been done explicitly by changing the annotation in the figure to be 'Asking questions of the data'. In addition to this, the same phrase has been used in the text when describing the process that it describes in Figure 1.

> Add in the and of section 1, add a new paragraph that presents the organization of the paper.

As can be seen at the end of the 'Introduction' section, there is now a summary of the organisation of the paper by section.

> Unify the symbols such as Figure or Fig.

This was a minor oversight that has now been amended. Thank you for bringing this to our attention.

> The related works section is not provided. I suggest to increase the number of studies and add anew discussion there to show the advantage, disadvantage, and weakness of the studied works. Authors should discuss the literature review more deep and clearly.

We have expanded the size of the bibliography in the paper, adding references to enrich the introduction and algorithm description. These are as follows:

- Abualigah, L.M., Khader, A.T., Hanandeh, E.S.: A combination of objective functions and hybrid krill herd algorithm for text document clustering analysis. Engineering Applications of Artificial Intelligence 73, 111 - 125 (2018). DOI 10.1016/j.engappai.2018.05.003
- Abualigah, L.M., Khader,A.T., Hanandeh,E.S.: Hybrid clustering analysis using improved krill herd algorithm. Applied Intelligence 48(11), 4047 - 4071 (2018). DOI 10.1007/s10489-018-1190-6
- Campos, G., Zimek, A., Sander, J., Campello, R., Micenkov, B., Schubert, E., Assent, I., Houle, M.: On the evaluation of unsupervised outlier detection: measures, datasets, and an empirical study. Data Mining and Knowledge Discovery 30(4), 891 - 927 (2016). DOI 10.1007/s10618-015-0444-8
- Chen, Y., Elliot, M., Smith, D.: The application of genetic algorithms to data synthesis: A comparison of three crossover methods. In: Privacy in Statistical Databases, pp. 160 - 171. Springer International Publishing (2018). DOI 10.1007/ 978-3-319-99771-1 11
- Dau, H.A., Keogh, E., Kamgar, K., Yeh, C.C.M., Zhu, Y., Gharghabi, S., Ratanamahatana, C.A., Yanping, Hu, B., Begum, N., Bagnall, A., Mueen, A., Batista, G., Hexagon-ML: The UCR time series classification archive (2018). https://www.cs.ucr.edu/~eamonn/time_series_data_2018/

- Liu, L., Cheng, L., Liu, Y., Jia, Y., Rosenblum, D.: Recognizing complex activities by a probabilistic interval-based model (2016)
- Olson, R.S., La Cava, W., Orzechowski, P., Urbanowicz, R.J., Moore, J.H.: PMLB: a large benchmark suite for machine learning evaluation and comparison. BioData Mining 10(1), 36 (2017). DOI 10.1186/s13040-017-0154-4
- Vikhar, P.A.: Evolutionary algorithms: A critical review and its future prospects. In: 2016 International Conference on Global Trends in Signal Processing, Information Computing and Communication (ICGTSPICC), pp. 261265 (2016). DOI 10.1109/ICGTSPICC.2016.7955308
- Wang, N., Shi, J., Yeung, D.Y., Jia, J.: Understanding and diagnosing visual tracking systems (2015). DOI 10.1109/ICCV.2015.355
- Zhao, W., Ma, H., He, Q.: Parallel k-means clustering based on MapReduce. In: Cloud Computing, pp. 674 - 679. Springer Berlin Heidelberg (2009). DOI 10.1007/978-3-642-10665-1 71

As is discussed throughout this response, the proposed method has no contemporaries in this paradigm. As such, the close discussion of the studied works would be not of benefit to the paper as they are not relevant in the classical sense where their contents is comparable to that of this work. The purpose of the introduction, really, is to provide the motivation of the paper of which the proposed method forms only a part. The root of the motivation (summarised in the numbered list in the introduction) is not novel and is addressed in a number of the cited works in the introduction. The fact that the proposed method is an example of a method from this paradigm is discussed in the later part of the introduction where other methods are addressed.

We refer to the editor if a standalone 'related works' section is required since it is not included as part of the Submission Guidelines. The guidelines are available at the following link: `https://www.springer.com/journal/10489/submission-guidelines?IFA`.

> Most references in the list of reference are very old, therefore, must update the reference list to contain articles related off at least five years and indexing in ISI and Scopus Database, in general update that list by the following reference related to predictions: 1-Abualigah, L. M. Q. (2019). Feature Selection and Enhanced Krill Herd Algorithm for Text Document Clustering. Studies in Computational Intelligence. 2-Abualigah, L. M. Q., & Hanandeh, E. S. (2015). Applying genetic algorithms to information retrieval using vector space model. International Journal of Computer Science, Engineering and Applications, 5(1), 19. 3-Abualigah, L. M., & Khader, A. T. (2017). Unsupervised text feature selection technique based on hybrid particle swarm optimization algorithm with genetic operators for the text clustering. The Journal of Supercomputing, 73(11), 4773-4795. 4-Abualigah, L. M., Khader, A. T., & Hanandeh, E. S. (2018). Hybrid clustering analysis using improved krill herd algorithm. Applied Intelligence. 5-Abualigah, L. M., Khader, A. T., & Hanandeh, E. S. (2018). A Combination of Objective Functions and Hybrid Krill Herd Algorithm for Text Document Clustering Analysis. Engineering Applications of Artificial Intelligence. 6-Abualigah, L. M., Khader, A. T., & Hanandeh, E. S. (2017). A new feature selection method to improve the document clustering using particle swarm optimization algorithm. Journal of Computational Science.

While some of the references are older than 5 years, this is often due to the fact that they are seminal papers for particular concepts, algorithms and software packages. For this reason, they are entirely relevant to the paper and have been included.

Taking this comment into account, however, we have revised our reference list to include two of the suggested articles (4 & 5) and other contemporary works where appropriate.

```
You need to explain clearly your proposed methods epically for the proposed
method.

Add a new figure to show the general procedures of the proposed method
```

We believe that the proposed method is explained in great detail with supplementary diagrams and algorithms for each subprocess; to add an 'epic' explanation to this description would make the manuscript cumbersome in terms of its length.

In terms of adding a new figure, there are diagrams which describe the concept and motivation of the work (Figure 1), the general structure of the method algorithmically (Figure 2), and each subprocess of the method (Figures 3-6). It is unclear what kind of figure is being asked for here which has not already been covered.

We refer to the editor on these points as the description of the method has been a point of pride for the authors that has been specifically commended when the work has been presented.

```
For the experimental results, it will be good to present a statistical test
in the comparison of the results with other published methods. This can help
to support the claim on improved results obtained with the selection methods
studied.
```

As is discussed in the introduction, the proposed method comes from a novel paradigm in which there are no 'other published methods' to compare it with. In this paradigm, the objective of the method is to generate data rather than to complete some comparable task X where one would be able to do some analysis like 'method A on X vs. method B on X'. As this is not possible, there is no sensible grounding to use a statistical test in this case.

```
What are the pros and cons of the proposed method? Please respond to this
question in the article text.
```

The advantages and disadvantages of the proposed method are now discussed more explicitly in the conclusion of the paper. In particular, issues around the total coverage, premature termination and simplistic solutions are discussed with reference to Figure 15.

```
in this paper, as you claimed, benchmark datasets are proposed. This is not
clear in the paper
```

This paper does not propose benchmark datasets. Where benchmark datasets are mentioned, it is to motivate the proposed paradigm and method, and not for comparison. Within this discussion, references are provided to articles that use benchmark datasets and to compilations of benchmark datasets themselves. We have clarified this in the text.

Reviewer 2

> The key concept behind the work is the evolutionary generation of datasets in order to understand the limitations and capabilities of ML algorithms. Thus, for this to be effective, you need to demonstrate that your representation and genetic operators are capable of permitting evolution to generate any possible dataset - or at least the set of datasets that would adequately cover the sets corresponding to the ML algorithm under investigation. e.g. If K-means needs to be investigated with sets A,B,C,D,E (each of which might be a different class of distribution) but your evolutionary approach is only capable of generating sets A,C,E, then you may gain an incomplete or misleading view of the capabilities of the algorithm. It is a big ask perhaps to prove that your method can generate *any* data distribution (or is not overly biassed towards the generation of some compared to others because of the representation or operators) and I'm sure later versions would be able to improve in these areas, but this issue is so important for the approach to be viable, that I think you should provide some evidence in the paper that the method can produce adequate coverage.

Thank you for this extensive comment. We agree that it may be an overstatement to say that our method can produce **any** dataset so we have revised our language throughout the text. More importantly, though, we have addressed the ability for our method to produce 'adequate' coverage. In Figure 15, there is a scatter plot showing the distribution of all individuals from every 50 generations using the first of the DBSCAN examples. The main conclusions of this plot are discussed in the conclusion of the article but may summarised as: (a) the method has covered a very substantial part of the search space, and (b) the method is indeed able to identify some preferable behaviour rather than persisting a full random cloud over the unit square.

> The second concern is the way evolutionary algorithms tend to find the easiest way out - so how can you stop the EA from evolving the simplest, easiest datasets for each ML algorithm (or the most difficult datasets if you reverse the fitness), instead of exploring the range of possibilities? (One option is to move more into curiosity-driven search, perhaps?)

This comment has been addressed by quite a substantial expansion of the conclusion. In the last few paragraphs, a meta-learning methodology is suggested for how this pitfall may be avoided within the scope of the proposed method. In this discussion, we address the fact that although the method is unsupervised, by designing an appropriately sophisticated fitness function a user could adopt a reinforced (or otherwise semi-supervised) learning behaviour with the proposed method.

We have also stressed (in the introduction and conclusion) the particular reasons for choosing an EA in this work, and that this method is just an example of what may come from further research within this new paradigm.