December 20, 2018

To the Editor:

Please find enclosed the revised version of our jointly authored paper “Inducing Non-Orthogonal and Non-Linear Decision Boundaries in Decision Trees via Interactive Basis Functions” that you kindly invited us to revise and resubmit for publication consideration to Expert Systems with Applications.

We would like again to thank two anonymous reviewers for their feedback on the revised version of this paper. Their suggestions for revisions were mostly to present some additional results and clarify some choices for the benchmarking experiment.

In this letter we describe our response to these suggestions, and the actions taken to address them. We trust that you will find that our actions provide a meticulous and conscientious response to the recommendations from the reviewers.

We thank you again for the attention paid to this submission and look forward to hearing back from you in due course.

Sincerely,

Antonio Páez

(on behalf of the coauthors)

**Reviewer #1 (responses in blue)**

In my opinion, in order to improve the paper, the authors can present accuracy values for each data set and for each classifier (IBFs) in table(s). Also, for Statistical Test, they can perform the Friedman test followed by Bonferroni-Dunn test (suited when comparing among multiple classifiers over multiple data sets).

Thank you for these valuable suggestions.

As recommended, we have a new table (Table 3) that presents the accuracy for each data set, algorithm (tree, forest, evolutionary tree), and implementation (orthogonal and IBF).

The data on Table 3 were used to perform the Friedman test. This we did in two ways. First, we performed the test on the whole table – and the results are significative. Based on our previous results, we expected that this would be due to the choice of algorithm (see Table 2 and Figure 8). So, we followed up with comparisons between orthogonal and IBFs *within* each algorithm (note that since this is a pairwise comparison a Bonferroni adjustment for multiple comparisons is not needed). The results of these tests were NOT significant, and we fail to reject the null hypothesis that the “treatments” (orthogonal and IBF) are not different.

This conclusion is not incorrect. In Table 2, for example, it can be seen there that on average IBFs perform only slightly better than orthogonal partitions. Figures 9, 10, and 11, on the other hand, show that IBFs perform better in around 50% of the datasets. Not surprisingly, the test does not detect a difference.

On the other hand, an important limitation of a univariate statistic such as Friedman’s is that it does not accounts for the attributes of the datasets. To give a Wikipedia example (https://en.wikipedia.org/wiki/Friedman\_test), the Friedman test is used as follows:

Suppose that “n welders each use k welding torches, and the ensuing welds were rated on quality. Do any of the k torches produce consistently better or worse welds?”

This assumes that the welds are performed on a similar material – no variations there. But if the welds were used on different materials (for instance, a selection of metals with different melting points), it is possible that the results might change, for instance if some torches are better suited to work with materials that have lower melting points.

The analog to this is the wide variation in the characteristics of the datasets, both in terms of their number of observations and compositions (number of features, number of classes, proportion of majority class). The Friedman test obscures the fact that orthogonal partitions might be more suitable in some conditions but not in others. Furthermore, since the Friedman test is based on the ranks and not the magnitudes of the differences, we also have no way to know how much better each treatment is with respect to the alternative. As seen in Figures 9, 10, and 11, IBFs perform better than orthogonal partitions only about half of the time. However, when they do perform better, the performance gains can be quite substantial.

For these reasons, a more refined way to understanding whether this is the case is a multivariate approach. This is done in two ways (one of which was already in the previous version of the paper). First, we estimate a model using the accuracy of all results in the 4-fold experiment (this was previously reported and becomes Model 1 in the present version of the paper). And secondly, we estimate the model using, for each dataset/method/partition style, the best performer of the 4 folds (Model 2). This second reinforces the findings reported in the previous version of the paper.

**Reviewer #2**

Why there are errors in the other ~30 datasets? Have you investigated the reasons?

We appreciate this question. The issues were mostly related to large datasets and the evolutionary tree algorithm, which may be a limitation of our computer resources rather than the algorithm.

This is explained in the revised version of the paper thusly:

“After excluding a number of datasets that led to errors with the evolutionary tree algorithm we tested, we work with the 93 datasets listed in Table 1. The main reason for the errors was a large number of observations, which may be a limitation of our computer resources rather than the algorithm.”