# Reply letter to second round of review

Marc Claesen Frank De Smet Johan A.K. Suykens Bart De Moor

October 17, 2014

We would like to thank all reviewers and the editors for their valuable comments and suggestions related to our submitted manuscript. In this document we will briefly summarize the modifications and extensions we made to the manuscript to address the comments we received.

The main improvements to the manuscript are as follows:

- We have doubled the number of iterations of each experiment from 10 to 20.
- We have added a brief discussion regarding the number of iterations in Section 5.4.
- We added experiments on another data set (`covtype`).
- We have added critical difference diagrams [1] to visualize the performance of each approach in Section 5.5.1 (class-weighted SVM, bagging SVM and RESVM) across all data sets. This provides increased insight in the performance of each approach across all data sets per setting.

## 1 Reviewer comments

**I would add more datasets in the experiment, but that is only a mild suggestion.**

We have added experiments on the `covtype` data set in each setting [2].

**I do not agree with your reply to my first comment about the number of repetitions in the experiments. The reply says: "Having ...more iterations...would favor RESVM...by yielding more significant outcomes." I do not quite understand this statement. Do you mean that by carrying out less repetitions in your experiments the results are more reliable? If this is what you mean then I do not agree.**

There appears to be a misunderstanding of our previous reply on this issue. The $p$-values that are computed in this manuscript depend on two factors, namely the true effect size (e.g. consistent difference in ranks of area under PR curve, a larger difference will yield lower $p$-values) and the sample size (e.g. number of repetitions, more repetitions will yield lower $p$-values for a given effect size). If we increase the number of repetitions, very small (consistent) performance improvements will induce statistically significant test results even when the improvements are of no practical relevance (e.g. too small). This discussion was included in the manuscript (Section 5.4). We agree with the reviewer that this only holds when the improvements are consistent and the difference between the previous and current revision of the manuscript attest to the consistency.

We have doubled the number of iterations in all experiments (10 to 20). The effect of obtaining more significant test results is already apparent in the PU learning setting on `mnist` and the semi-supervised setting (all data sets except `covtype`). RESVM now obtains much more statitistically significant improvements over

bagging SVM compared to the previous revision, even though the actual improvement in performance (*effect size*) remains the same. The updated manuscript indicates the performance improvements are consistent (illustrated by an increased amount of significant results). Given the consistency in both confidence intervals and win counts of the learning approaches, we believe that increasing the number of repetitions further would only increase the number of statistically significant test results (e.g. make RESVM look better in comparison) in the situations where RESVM is marginally better than bagging SVM.

**Averages with less repetitions are more variable and hence could yield to significant results when they are not, in favor of one or another algorithm.**

The statistical tests are arranged to control the Type I error rate (to $\alpha = 0.05$), regardless of sample size (incorrectly rejecting the null hypothesis of no difference between bagging SVM and RESVM is a Type I error). In small sample size situations (few repetitions), these tests require larger effect sizes to yield a significant outcome. This means that a true performance improvement is more likely to be missed due to insufficient power (e.g. insignificant test result) but not the other way around. This is why we noted that increasing the number of repetitions would favor RESVM as it would increase the test's power while keeping the Type I error rate constant. This will lead to *more* statistically significant results when the number of repetitions is raised. The amount of significant results in this revision attest to this observation.

**I still think that 100 repetitions would be better.**
The primary goal of our manuscript is to provide evidence of a *practically relevant* performance improvement over existing approaches. We believe that increasing the number of repetitions further is not warranted (after doubling the repetitions in this revision). As an illustration, we ran 100 repetitions on the `covtype` data set to compare it to using 20 repetitions in Section 5.4 which highlights that the only change is an additional statistically significant test result.

# References


[1] Janez Demšar. Statistical comparisons of classifiers over multiple data sets. *Journal of Machine Learning Research*, 7:1–30, 2006.

[2] Jock A. Blackard and Denis J. Dean. Comparative accuracies of artificial neural networks and discriminant analysis in predicting forest cover types from cartographic variables. *Computers and Electronics in Agriculture*, 24(3):131–151, December 1999.