Prof. Seokho Lee
Department of Statistics
Hankuk University of Foreign Studies
81 Oedae-ro, Yongin, Gyeonggi, South Korea 449-791 July 7, 2016

Editor
Communications for Statistical Applications and Methods

Dear Editor,

Thank you for your letter of July 3, 2016 inviting us for a minor revision of our paper entitled "Ensemble approach for improving prediction in kernel regression and classification," (CSAM-2016-094). Enclosed please find our revision and item-by-item responses to the comments from the two reviewers. We have carried out a revision of the paper and believe that the current revision addresses all comments on the previous submission.

Thank you very much for your consideration. We look forward to hearing from you.

Yours sincerely,

Seokho Lee

Response to the Comments of Reviewer A

We are grateful for the reviewer's comments, which have helped us significantly improve the paper. In what follows, we state each of the comments in italics, and describe our response in plain text.

1. Brief review in Section 2.2 is too brief. I hope the authors write or survey more contents or recent development including crucial contents regarding ensemble approaches.

Ensemble methods have become a major learning paradigm since 1990's. They comprise of bagging, random forests, boosting, and various variants of them. Ensemble methods have been widely studied in various research areas, including statistical learning, machine learning, pattern recognition, computational intelligence, robotics, artificial intelligence, game theory, neural networks, to name a few. They have been extensively applied in vast quantitative research areas, including bioinformatics, cognitive science, natural language processing, computer vision, to name a few. It is very challenging and out of an aim of this paper to provide a comprehensive, or even selective, survey on recent development of ensemble method in this manuscript. Instead, we introduce a couple of references (Bühlmann, 2012; Zhou, 2012) in the revised manuscript in order to compensate the lack of the full review on ensemble approach. We have added the paragraph in Section 2.2 of the revised manuscript that "Ensemble methods have become a major learning paradigm since 1990's. This learning approach comprises of a lot of methodologies, including bagging, random forests, boosting and various variants of them. They have been extensively applied in vast application areas, including bioinformatics, cognitive science, natural language processing, computer vision, to name a few. An extensive research on ensemble learning and its numerous applications can be found in Bühlmann (2012), Zhou (2012), and references therein."

2. Please give any references on OOB if there are.

We have added a standard reference (Breiman, 1996) for OOB in the revised manuscript.

References

Breiman, L. (1996), *Out-of-bag estimation*, Technical report, Department of Statistics, University of California at Berkeley, CA, USA.

Bühlmann, P. (2012) Bagging, boosting and ensemble methods. In Gentle, J. E., Härdle, W. and Mori, Y. (eds), *Handbook of Computational Statistics*. Springer-Verlag, 877–907.

Zhou, Z. (2012). *Ensemble Methods: Foundations and Algorithms*, CRC Press, Boca Raton.

Response to the Comments of Reviewer B

We are grateful for the reviewer's comments, which have helped us significantly improve the paper. In what follows, we state each of the comments in italics, and describe our response in plain text.

It is a good paper to be publishable in CSAM.

Thank you very much for your kind comment. We have addressed an issue you raised, see our response below.

I would be happy to see one or two examples to show how the ensemble approach really useful in a real dataset, in addition to just the evaluation test using RMSE or misclassification rate. I mean that may include the meaning or interpretation of the ensemble predictor for real example, advantage or disadvantage in real data analysis, and comparison to other approaches.

A main strength of ensemble approaches is to provide a final learner of increased prediction ability. But the price to pay for prediction enhancement is loosing interpretational ease of the final learner. This can be easily understood that, in tree models, a single tree has a direct interpretation for terminal nodes while an ensemble tree does not maintain a tree structure anymore. This kind of pros and cons of ensemble methods can be found in any standard textbook on machine learning, for example Hastie et al. (2011). Thus, when we invoke bagging or random forests even in the parametric models, we are not able to attempt the interpretation on the final model in terms of predictors unless sub-learners are linear in predictors. (But one of heuristic ways of measuring the relative importance of predictors in the ensemble model can be found in Hastie et al., 2011.)

Especially, kernel-based models in our manuscript deal with nonlinear relation between predictors and the response. The mean of response, or a function of it, is assumed to be linearly expressed with basis of the feature space, induced from the kernel function used, whose form is typically not known a priori nor even tried to be inferred.

To sum up, it is challenging for us to provide an example where *the meaning or interpretation of the ensemble predictors* other than prediction

performance. The latter is, we believe, well described in the manuscript. We would like to ask for your understanding.

References

Hastie, T., Tibshirani, R., and Friedman, J. (2011). *The Elements of Statistical Learning, 2nd Edn.*, Springer, New York.

Response to the Associate Editor

We are grateful for the A.E.'s comments, which have helped us significantly improve the paper. In what follows, we state each of the comments in italics, and describe our response in plain text.

Upon minor revision, this ms is acceptable for publish in CSAM. I have some recommendation for the authors, in addition to the comments by two reviewers.

Thank you very much for your kind comment. We have addressed all issues you listed, see our response below.

1. Please mention on statistical software for this paper, and hopely the authors provide it to CSAM repository as a supplementary material.

Kernel ridge regression (KRR) and kernel logistic regression (KLR) are optimized straightforwardly. The solution of KRR has a closed-form solution which is a ridge-type least squares solution. However KLR resorts iteratively reweigted least squares method under ridge-type penalty. The updating formula at each iteration is given a ridge-type weighted least square solution. We implemented them using a statistical software `R`. Ensemble methods, bagging and random forest in our paper, fit these models for multiple bootstrap samples. Thus, remaining works we did are drawing multiple bootstrap samples using a sampling function (`sample()` in R, for example), fitting KRR and KLR to each bootstrap sample, and combining such bootstrap subfits for a final prediction.

All of the above procedures can be easily implemented using any statistical software or programming language, not necessarily requiring `R`. In our `R` implementation, we did not use any specific package. Thus, it does not worth mentioning the implementation details in the manuscript. However, to make it more specific, we have added a sentence in Section 2.1 of the revised manuscript that "For a given λ, the solution of (2.1) has a closed-form solution $\hat{\mathbf{d}}_\lambda = (\mathbf{K} + \lambda \mathbf{I}_n)^{-1}\mathbf{y}$ with $\mathbf{y} = (y_1, \ldots, y_n)^T$. The solution of (2.2) can be obtained through a standard iterative procedure using quadratic approximation (iteratively reweighted least squares), where the updating formula for $\mathbf{d}$ is given

as a solution of ridge-type weighted least squares. This can be easily implemented using any standard statistical software and we used a statistical software R for our implementation."

2. Please add more sentences in Abstract to make it at least 8 lines.

Taking your advice, we have changed the Abstract in the revised manuscript, which has 8 lines.

3. Please add more keywords to make it at least 7.

Taking your advice, we have additionally added in the revised manuscript 2 more keywords "bootstrap, ensemble method" in the list of keywords, resulting in 7 keywords in total.

4. Some presentation in section 2.3 are not so clear. Hope the authors provide more details on the contents in section 2.3 which is the most important part in this ms.

We believe that the core of our approach is described well in Section 2.3 of the original manuscript. However, we admit that simply providing the plain paragraph is not helpful for potential readers' understanding. To deliver a clear idea of the procedure, we provide the conceptual algorithm in Section 2.3 of the revised manuscript as below:

Algorithm 1: Ensemble method in kernel-based regression and classification

1. Set $\sigma = 1/p$ (bagging) or $\sigma = 1/m$ (random forest)
2. For $b = 1, \ldots, B$ and for λ on a pre-specified grid, repeat
   (a) Create a bootstrap sample $\mathcal{D}_b$, and set an OOB sample $\mathcal{O}_b$ consisting of the observations not included in $\mathcal{D}_b$. In this step, $\mathcal{D}_b$ and $\mathcal{O}_b$ contains randomly selected $m \approx \sqrt{p}$ predictors for random forest, $m = p$ predictors for bagging.
   (b) Compute the kernel matrix $\mathbf{K}_b$ on $\mathcal{D}_b$, and calculate $\hat{\mathbf{d}}_\lambda^b$.

(c) Compute the kernel matrix $\mathbf{K}_b^*$ between $\mathcal{D}_b$ and $\mathcal{O}_b$, and compute $\hat{f}_{b,\lambda}(\mathbf{x}_i)$ for $\mathbf{x}_i \in \mathcal{O}_b$.

(d) Compute an OOB error, say $\mathrm{OE}_b(\lambda)$, by applying $\hat{f}_{b,\lambda}(\mathbf{x}_i)$ for $\mathbf{x}_i \in \mathcal{O}_b$.

3. Compute $\mathrm{OE}(\lambda) = \frac{1}{B}\sum_{b=1}^{B} \mathrm{OE}_b(\lambda)$, and find the optimal λ as $\lambda_{opt} = \arg\min_\lambda \mathrm{OE}(\lambda)$.

4. Given a new observation $\mathbf{x}_*$, for $b = 1, \ldots, B$, repeat

(a) Obtain $\hat{\mathbf{d}}^b_{\lambda_{opt}}$ using $\mathbf{K}_b$ and $\mathcal{D}_b$ used in Step 2.

(b) Compute $\mathbf{K}_b(\mathbf{x}_*) = (K(\mathbf{x}_1, \mathbf{x}_*), \ldots, K(\mathbf{x}_n, \mathbf{x}_*))^T$ and compute $\hat{f}_{b,\lambda_{opt}}(\mathbf{x}_*)$.

(c) Compute $\hat{f}_{\lambda_{opt}}(\mathbf{x}_*) = \frac{1}{B}\sum_{b=1}^{B} \hat{f}_{b,\lambda_{opt}}(\mathbf{x}_*)$

5. Report $\hat{y} = \hat{f}_{\lambda_{opt}}(\mathbf{x}_*)$ for regression and $\hat{y} = \{\mathrm{sign}(\hat{f}_{\lambda_{opt}}(\mathbf{x}_*)) + 1\}/2$ for classification as the prediction for $\mathbf{x}_*$.
