Discussion of Random Projection Ensemble Classification by Timothy I. Cannings and Richard J. Samworth

Yining Chen

London School of Economics and Political Science, London, U.K.

Rajen D. Shah

University of Cambridge, Cambridge, U.K.

We congratulate the authors for this interesting paper which introduces an important ensemble method for random projections in classification problems. We shall limit our comments to the procedure of selecting random projections and aggregating the results.

The basic procedure, as stated in Section 3, involves forming $B := B_1 \times B_2$ random projections of the data. A base classifier (e.g. k-nearest neighbours) is trained on each of these B projected versions of the data. The resulting classifiers are then grouped consecutively into blocks of size B_2, where we pick and then average the ones with the lowest training or leave-one-out (LOO) cross-validation error from each group, and discard the rest. However, the blocking strategy perhaps does not make full use of the information from the training or LOO estimates whose construction is usually the most computationally intensive part of the procedure. Indeed, grouping base classifiers consecutively is somewhat arbitrary: the distribution of the ensemble classifier, conditional on the data and the set of random projections, is unchanged when permuting the list of classifiers. Therefore, one can construct new ensemble classifiers resulting from multiple random groupings with little extra computational cost. Here each new classifier is still based on the B base classifiers, but we instead randomly permute the order of the base classifiers before grouping them into blocks consecutively. By aggregating these new classifiers by a simple majority vote, we form a final classifier, which could potentially remove some of the variance resulting from the randomness of the grouping.

To examine the performance, we applied both the original method and the variant to four real datasets using k-nearest neighbours with different training set sizes and setting $B = 1000$ and $B_2 = 50$. Results are reported in Figure 1. As expected, the proposed variant with multiple random grouping gives slightly improved performance.

More generally, we could think of the training/LOO predictions as new training data for a further classifier; an approach known as stacking or blending (Wolpert, 1992; Breiman, 1996). We looked at forming a final classifier via regression of the class labels on the LOO predictions of k-nearest neighbours using ℓ_1-penalised logistic regression with a non-negativity constraint on the coefficients. This can be viewed as a data-driven way of forming a weighted average of B classifiers on the projected versions of the data. Results on the eye state data (see Section 6.2.1, where RP-$k$nn$_5$ performed the best) with $n = 1000$ are shown in Table 1. This suggests that some slightly more data-driven variants of the aggregation procedure used in the paper may lead to further improved performance in some settings, even with a smaller B.

References

Breiman, L. (1996). Stacked regressions. *Machine Learning*, **24**, 49–64.

Wolpert, D. (1992). Stacked generalization. *Neural Networks*, **5**,241–259.

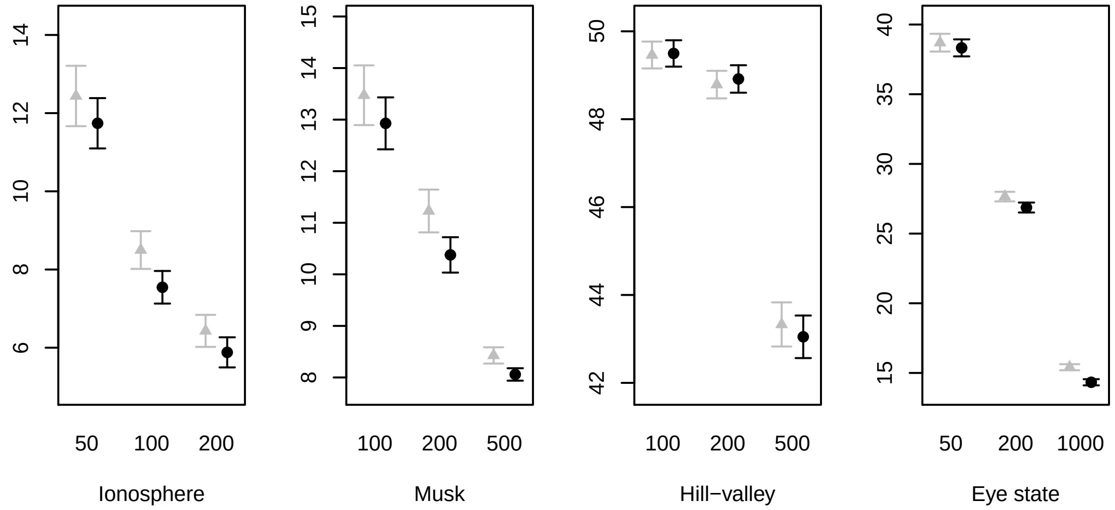

Fig. 1. Misclassification rates and the corresponding confidence intervals of the original random projection ensemble classifier (grey) and the multiple random grouping approach (black) on four real datasets (considered by the authors in Section 6.2) with different training set sizes and $(B, B_2) = (1000, 50)$.

Table 1. Estimated misclassification rates and the corresponding standard errors of different classifiers for the eye state data

Classifier	Misclassification rate
k-nn	$14.45_{0.16}$
RP-$k\text{nn}_5, B = 25000$	$13.54_{0.19}$
RP-$k\text{nn}_5$ with stacking, $B = 500$	$12.86_{0.08}$
RP-$k\text{nn}_5$ with stacking, $B = 5000$	$11.35_{0.07}$