Mixing hetero- and homogeneous models in weighted ensembles

James Large and Anthony Bagnall

University of East Anglia, Norwich Research Park, UK
{james.large, ajb}@uea.ac.uk

Abstract. The effectiveness of ensembling for improving classification performance is well documented. Broadly speaking, ensemble design can be expressed as a spectrum where at one end a set of heterogeneous classifiers model the same data, and at the other homogeneous models derived from the same classification algorithm are diversified through data manipulation. The cross-validation accuracy weighted probabilistic ensemble is a heterogeneous weighted ensemble scheme that needs reliable estimates of error from its base classifiers. It estimates error through a cross-validation process, and raises the estimates to a power to accentuate differences. We study the effects of maintaining all models trained during cross-validation on the final ensemble's predictive performance, and the base model's and resulting ensembles' variance and robustness across datasets and resamples. We find that augmenting the ensemble through the retention of all models trained provides a consistent and significant improvement, despite reductions in the reliability of the base models' performance estimates.

Keywords: classification, ensembles, heterogeneous, homogeneous

1 Introduction

Broadly speaking, there are three families of algorithms that could claim to be state of the art in classification: support vector machines; multilayer perceptrons/deep learning; and tree-based ensembles. Each has their own strengths on different problem types under different scenarios and contexts. However, our primary interest is when faced with a new problem with limited or no domain knowledge, what classifier should be used?

[10] introduced the Cross-validation Accuracy Weighted Probabilistic Ensemble (CAWPE), a weighted ensemble scheme [9] which demonstrates consistent significant improvements over its members, significant improvements over competing heterogeneous ensemble schemes, and, at worst, no significant difference to large homogeneous ensembles and heavily tuned state of the art classifiers. The key premise is that given a lack of domain knowledge to suggest one particular type of model over another, the best place to start on a new arbitrary problem is to heterogeneously ensemble over different kinds of classifiers instead of invest into a single type. This idea is nothing new of course, however a far larger share of the

literature has typically been devoted to homogeneous ensembling or optimisation of individual models.

CAWPE cross-validates each member on the train data, generating error estimates, and then raises these estimates to the power $\alpha = 4$ [10] to generate accentuated weightings for the members' probability estimates when predicting. Models rebuilt on the full train data are used to form predictions for the ensemble. To this end, it needs to gather reliable error estimates. We do this through a ten-fold cross-validation (CV) [8]. During the CV process, however, models are made on each fold which are then discarded. A natural question is whether these can be retained and leveraged to improve predictive performance.

We investigate whether retaining these models, in addition to the models retrained to the full training set, can improve classification performance. We also assess whether accuracy can be maintained while skipping the retraining step on the full data, saving time in the training phase. While maintaining these models incurs no additional training time cost, prediction time and space requirements clearly increase in proportion to the number of CV folds. We further analyse the variance of the maintained classifiers and their effects on the resulting ensemble's variance.

Explicitly building homogeneous (sub-)ensembles from heterogeneous base classifiers is not a new idea. [7] builds forests of trees from different tree building algorithms and shows that larger purely-homogeneous forests can be matched or beaten by smaller mixed forests. Ensemble selection [4] (or pruning) can similarly be applied to purely hetero- or homogeneously generated model sets, or mixtures of the two [11]. Alongside these works, we specifically wish to study the effects of maintaining homogeneous models, with potentially lower-quality estimates of competency attached, on the CAWPE weighting scheme which relies heavily on the weightings applied.

We outline our experimental procedure in Section 2. Results are summarised in Section 3, and analysed further in Section 4. We conclude in Section 5.

2 Experimental Setup

The UCI dataset archive[1] is widely used in the machine learning literature. We have taken 39 real-valued, independent and non-toy datasets to use in our experiments, following feedback received on the superset of these datasets used in [10]. The datasets are summarised in Table 1.

Experiments are conducted by averaging over 30 stratified resamples of each dataset. Data, results and code can all be found at the accompanying website for this research[2]. For each resample, 50% of the data is taken for training (on which the cross validation process for each base classifier is performed), 50% for testing. We always compare classifiers on the same resamples, and these can be exactly reproduced with the published code.

[1] http://archive.ics.uci.edu/ml/index.php
[2] http://www.timeseriesclassification.com/CAWPEFolds.php

Table 1. A full list of the 39 UCI datasets used in our experiments. Full names saved for horizontal space: *[1] conn-bench-sonar-mines-rocks, *[2] conn-bench-vowel-deterding, *[3] vertebral-column-3clases.

Dataset	#Cases	#Atts	#Classes	Dataset	#Cases	#Atts	#Classes
bank	4521	16	2	page-blocks	5473	10	5
blood	748	4	2	parkinsons	195	22	2
breast-cancer-w-diag	569	30	2	pendigits	10992	16	10
breast-tissue	106	9	6	planning	182	12	2
cardio-10clases	2126	21	10	post-operative	90	8	3
sonar-mines-rocks*[1]	208	60	2	ringnorm	7400	20	2
vowel-deterding*[2]	990	11	11	seeds	210	7	3
ecoli	336	7	8	spambase	4601	57	2
glass	214	9	6	statlog-landsat	6435	36	6
hill-valley	1212	100	2	statlog-shuttle	58000	9	7
image-segmentation	2310	18	7	statlog-vehicle	846	18	4
ionosphere	351	33	2	steel-plates	1941	27	7
iris	150	4	3	synthetic-control	600	60	6
libras	360	90	15	twonorm	7400	20	2
magic	19020	10	2	vertebral-column*[3]	310	6	3
miniboone	130064	50	2	wall-following	5456	24	4
oocytes_m_nucleus_4d	1022	41	2	waveform-noise	5000	40	3
oocytes_t_states_5b	912	32	3	wine-quality-white	4898	11	7
optical	5620	62	10	yeast	1484	8	10
ozone	2536	72	2				

We evaluate three ensemble configurations that retain the models evaluated on CV folds of the train data against the original CAWPE, which ensembles only over the models retrained on the entire train set. These are to a) (M)aintain all models trained on CV folds and add them to the ensemble alongside the fully trained models (CAWPE_M), b) (M)aintain all models once more, but systematically (D)own-(W)eight them relative to the fully trained models due to their potentially less reliable error estimates (CAWPE_M_DW) c) maintain *only* those models trained on the CV folds, and skip the retraining step on the full train data, (R)eplacing the original models (CAWPE_R).

All configurations of CAWPE tested use the same core base classifiers, those defined as the 'simple' set in [10] of logistic regression; C4.5 decision tree; linear support vector machine; nearest neighbour classifier; and a multilayer perceptron with a single hidden layer. These classifiers are each distinct in their method of modelling the data, and are approximately equivalent in performance on average. Because all dataset resamples and CV folds of the respective train splits are aligned, each ensemble configuration is therefore being built from identical (meta-)information and we are only testing the configuration's ability to combine the predictions.

For reference, we also compare to Random Forest (RandF) and eXtreme Gradient Boosting (XGBoost), each with 500 trees. This is to put the results into context, rather than to claim superiority or inferiority to them. XGBoost in particular would likely benefit from tuning, for example, which we do not perform for these experiments.

When comparing multiple classifiers on multiple datasets, we follow the recommendation of Demšar [5] and use the Friedmann test to determine if there are any statistically significant differences in the rankings of the classifiers. However, following recent recommendations in [1] and [6], we have abandoned the Nemenyi post-hoc test originally used by [5] to form cliques (groups of classifiers

within which there is no significant difference in ranks). Instead, we compare all classifiers with pairwise Wilcoxon signed-rank tests, and form cliques using the Holm correction (which adjusts family-wise error less conservatively than a Bonferonni adjustment).

We assess classifier performance by four statistics of the predictions and the probability estimates. Predictive power is assessed by test set accuracy and balanced test set accuracy. The quality of the probability estimates is measured with the negative log likelihood (NLL). The ability to rank predictions is estimated by the area under the receiver operator characteristic curve (AUC). For multiclass problems, we calculate the AUC for each class and weight it by the class frequency in the train data, as recommended in [12].

3 Results

We summarise comparative results succinctly here in three forms: Figure 1 displays CAWPE configurations and reference homogeneous ensembles ordered by average ranks in accuracy along with cliques of significance formed; Table 2 details the average scores of all four evaluation metrics; and Table 3 details pairwise wins, draws and losses between the original and proposed CAWPE configurations.

Maintaining the individual fold classifiers significantly improves over the original CAWPE. Within the three proposed configurations there is very little difference in performance. This is largely to be expected since they are working from the same meta-information, with the exception of CAWPE_R, which replaces the fully re-trained models only with those trained during CV. This does mean that training time can seemingly be saved by avoiding this final retraining step without a tangible reduction in predictive performance.

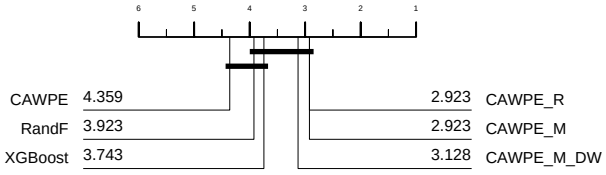


Fig. 1. Critical difference diagram displaying the average ranks of accuracy of the original CAWPE and three tested configurations and reference homogeneous ensembles. Classifiers connected by a solid bar are considered within the same clique and not significantly different from each other.

Note that while maintaining the fold classifiers improves performance with statistical significance, the average improvement in absolute terms is very small, roughly 0.3% in terms of accuracy, balanced accuracy, and area under the curve

Table 2. Averages scores for four evaluation metrics of each of the CAWPE configurations and homogeneous ensembles tested.

Classifier	ACC ↑	BALACC ↑	AUC ↑	NLL ↓
CAWPE	0.861	0.787	0.915	0.53
CAWPE_M_DW	0.864	0.789	0.917	0.517
CAWPE_M	0.865	0.79	0.918	0.515
CAWPE_R	0.865	0.789	0.918	0.516
RandF	0.854	0.78	0.91	0.564
XGBoost	0.85	0.784	0.907	0.647

(Table 2). Meanwhile, XGBoost's average accuracy is a full 1.2% lower, but still significantly similar to the new CAWPE configurations. This is because the improvement found while being small, is very consistent. When looking at the paired wins, draws and losses between the configurations in Table 3, the contrast between the relatively balanced match-ups of the three new configurations, against the consistently beaten original configuration is clear to see.

Table 3. Pairwise wins, draws and losses in terms of dataset accuracies between the ensemble configuration on the row against the configuration on the column.

	CAWPE_R	CAWPE_M	CAWPE_M_DW	CAWPE
CAWPE_R	-	17/4/18	23/0/16	32/0/7
CAWPE_M	18/4/17	-	23/0/16	31/0/8
CAWPE_M_DW	16/0/23	16/0/23	-	34/0/5
CAWPE	7/0/32	8/0/31	5/0/34	-

4 Analysis

CV is such a commonly used method of evaluating a model on a given dataset because of it's robustness and completeness relative to, for example, singular held-out validation sets [8]. A single fold of a CV procedure in isolation is of course simply the latter, and equivalent to a single subsample within a bagging context [2]; it is the repeated folding of the data that leads to each instance being predicted as a validation case once that makes the process complete.

All weighted ensembles rely to some extent on the reliability of the error estimates of their members, but CAWPE especially does given that it accentuates the differences in those estimates. We wish to analyse the extent to which the quality of error estimates suffers, and its effects on the ensemble's own performance and variance.

Figure 2 measures the counts of differences in estimated (on train data) and observed (on test data) accuracies and confirms expectations that completing the CV process and retraining models on the full dataset results in more accurate

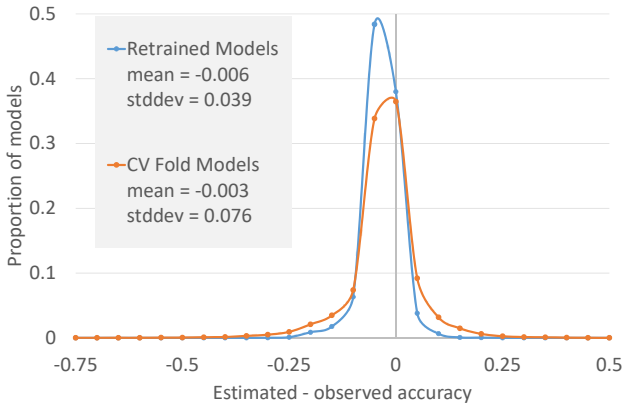


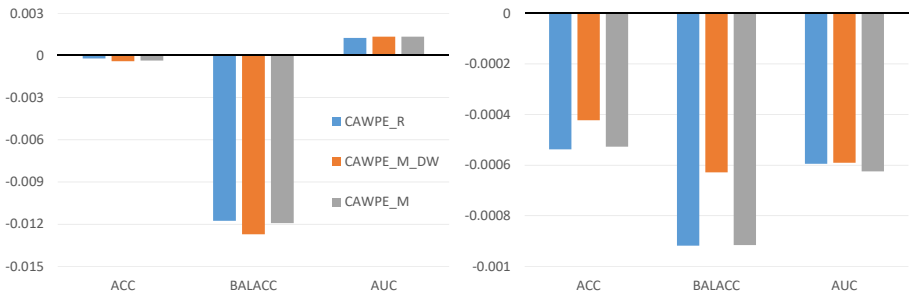
Fig. 2. Normalised counts of differences in estimated (on train data) and observed (on test data) accuracy for the retrained (blue) and individual CV fold (orange) models across all datasets and resamples. Positive x values indicate a larger estimated than observed accuracy, i.e. a classifier overestimating its performance.

estimates of accuracy on average than the individual models on CV folds. Overall standard deviation almost doubles, but the number and degree of the outliers is perhaps the most important thing. The retrained models never have performance under-estimated by more than 0.3, and less than 2% of the models under estimate by more than 0.1.

Meanwhile, the individuals fold estimates have some extreme outliers in terms of underestimating in particular, with a small tail on Figure 2 stretching all the way to -0.75. 7.6% of all fold models underestimate accuracy by more than 10%. Many of the extreme outliers were localised to two datasets, spread out across different learning algorithms. The breast-tissue dataset is a relatively balanced six class problem, while post-operative is a heavily imbalanced three class problem. These factors along with them being the datasets with the least instances likely lead to difficult folds to classify for certain models and seeds, which are of course averaged over when considering the remaining CV folds.

In context, however, the difference really is not too stark. The errors in estimates may double in variance, and these are being accentuated by CAWPE's combination scheme, but there are also fifty more models to average over. Figure 3 summarises the differences in variance across test performances between the configurations that maintain the fold models and the original CAWPE along two dimensions - variance in performance on arbitrary datasets, and variance in performance over formulations of the same dataset through resampling. Variance across resamples is reduced, while variance over datasets is less clear. It seems as though cases such as breast-tissue and post-operative affect this particular

comparison as with the above, and this shows with variance in balanced accuracy still being clearly reduced.



(a) Standard deviations of performances overs dataset

(b) Average standard deviations in performance over dataset resamples

Fig. 3. Standard deviations in performance metrics over (a) datasets and (b) dataset resamples of the three proposed CAWPE configurations, expressed as differences to the original. NLL is omitted due to the improper scaling factor brought about by it not being a measure in the range 0 to 1, however variance similarly drops for all three of the proposed configurations.

When there are only five members, erroneously discounting a classifier to the extent that it's outputs are effectively worthless is a large blow to the overall strength of the ensemble. In the case of ensembles with 50 or 55 members though, erroneously discounting one or two classifiers is not so harmful. Practitioners of homogeneous ensembles will of course be familiar with this, and it is the underpinning of the design of such an ensemble - averaging over high variance inputs to produce a low variance output [3].

5 Conclusions

We have experimentally evaluated the effectiveness of maintaining models used to estimate the accuracy of base classifiers in a weighted ensemble, in addition to or in place of the original models. The experiments show a minor but significant and very consistent improvement in performance across different evaluation metrics, even when skipping the retraining of the models on the full dataset. While variance in the estimates of performance that fuel the weightings within the ensemble increases, this is offset by the averaging effect of the greater number of models, as observed in typical homogeneous ensembles.

Further experimentation aims to discover a breaking point between the effectiveness of increased homogeneity versus heterogeneity. In these experiments,

as in previous, a ten fold cross-validation process was used to evaluate the models and ultimately as the source of the expanded model set. Increasing or decreasing the number of folds relative to time and space requirements, or switching entirely to a randomised bagging approach with heterogeneous members are interesting routes to follow.

Acknowledgement This work is supported by the UK Engineering and Physical Sciences Research Council (EPSRC) [grant number EP/M015807/1] and Biotechnology and Biological Sciences Research Council (BBSRC) Norwich Research Park Biosciences Doctoral Training Partnership [grant number BB/M011216/1]. The experiments were carried out on the High Performance Computing Cluster supported by the Research and Specialist Computing Support service at the University of East Anglia.

References

1. Benavoli, A., Corani, G., Mangili, F.: Should we really use post-hoc tests based on mean-ranks? Journal of Machine Learning Research **17**, 1–10 (2016)
2. Breiman, L.: Bagging predictors. Machine Learning **24**(2), 123–140 (1996)
3. Breiman, L.: Random forests. Machine Learning **45**(1), 5–32 (2001)
4. Caruana, R., Niculescu-Mizil, A.: Ensemble selection from libraries of models. In: Proc. of the 21st International Conference on Machine learning (2004)
5. Demšar, J.: Statistical comparisons of classifiers over multiple data sets. Journal of Machine Learning Research **7**, 1–30 (2006)
6. García, S., Herrera, F.: An extension on statistical comparisons of classifiers over multiple data sets for all pairwise comparisons. Journal of Machine Learning Research **9**, 2677–2694 (2008)
7. Gashler, M., Giraud-Carrier, C., Martinez, T.: Decision Tree Ensemble: Small Heterogeneous Is Better Than Large Homogeneous. 2008 Seventh International Conference on Machine Learning and Applications pp. 900–905 (2008). https://doi.org/10.1109/ICMLA.2008.154, `http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=4796917`
8. Kohavi, R.: A study of cross-validation and bootstrap for accuracy estimation and model selection. In: Proc. 14th International Joint Conference on Artificial Intelligence. pp. 1137–1143. Morgan Kaufmann Publishers Inc. (1995)
9. Kuncheva, L., Rodríguez, J.: A weighted voting framework for classifiers ensembles. Knowledge and Information Systems **38**(2), 259–275 (2014)
10. Large, J., Lines, J., Bagnall, A.: A probabilistic classifier ensemble weighting scheme based on cross-validated accuracy estimates. Data Mining and Knowledge Discovery (Jun 2019). https://doi.org/10.1007/s10618-019-00638-y, `https://doi.org/10.1007/s10618-019-00638-y`
11. Partalas, I., Tsoumakas, G., Vlahavas, I.: A study on greedy algorithms for ensemble pruning. Aristotle University of Thessaloniki, Thessaloniki, Greece (2012)
12. Provost, F., Domingos, P.: Tree induction for probability-based ranking. Machine Learning **52**(3), 199–215 (2003)