ESANN 2026 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 22-24 April 2026, i6doc.com publ., ISBN 9782875870964. Available from http://www.i6doc.com/en/.

Clagging: Generating and combining predictions using clustering

Arnaud Germain[1] and Frédéric Vrins[1] *

UCLouvain - LIDAM/Louvain Finance research center (LFIN)
Voie du Roman Pays, 34 - 1348 Louvain-la-Neuve - Belgium

Abstract. We introduce a new ensemble learning strategy called clagging (for *cluster aggregating*) which consists in combining models fitted on different clusters. First, we perform K clustering tasks on the same training set $\mathcal{D}$, increasing linearly the number of clusters from 1 to K. Next, we fit a model on each of those $1 + 2 + \ldots + K$ clusters. Finally, the output for a given test point is obtained by combining the predictions of the corresponding models using the distance of the test point to the clusters' centroids. We perform an extensive horse race study where we benchmark clagging on 10 regression datasets and 7 prediction algorithms. Our results suggest that clagging outperforms the standard version of bagging and typically performs best when choosing $K > 1$, indicating that it outperforms the considered model trained on the whole training set $\mathcal{D}$ ($K = 1$).

1 Introduction

Ensemble learning consists in fitting multiple models $\varphi_1, \ldots, \varphi_J$ (called *base learners* hereafter) and aggregating their predictions (*model combination*) to get a single output. This approach has been shown to outperform individual models $\hat{y} = \varphi(x)$ on a wide range of tasks, especially when there is some diversity in the set of base learners [1]. The purpose of ensemble learning is to reduce the variance of the prediction errors compared to the output of a single base learner φ fitted on the full training set $\mathcal{D}$. Several approaches can be considered to foster diversity in the set of base learners. A first class of approaches is to rely on different predictive algorithms φ_j (e.g., linear, non-linear, rule-based, tree-based, ...) fitted on the same training set $\mathcal{D}$ [2, 3]. Another class of approaches is to choose a model φ from a given family (e.g., a decision tree) and fit the latter on various subsets $\mathcal{D}_1, \ldots, \mathcal{D}_J$ of $\mathcal{D}$. The most popular example in this class is perhaps *bagging*, where the subsets $\mathcal{D}_j$ are bootstrapped samples drawn from $\mathcal{D}$. It has been shown to work both theoretically and empirically [4, 5]. In bagging, a bootstrapped sample $\mathcal{D}_j$ is usually created by drawing observations from the original training set $\mathcal{D}$ with replacement so that the size of the bootstrapped samples coincides with the size of the original training set, $\sharp\mathcal{D}_j = \sharp\mathcal{D}$. However, this choice is quite arbitrary and might be suboptimal in terms of generalization accuracy. For instance, it is shown empirically that the optimal choice for the

*The work of Arnaud Germain is funded by the ING chair. Frédéric Vrins benefits from the financial support of the Fonds de la Recherche Scientifique F.S.R.-FNRS (grant J.0225.24) as well as of the Belgian Federal Science Policy Office (grant ARC 18-23/089).

ESANN 2026 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 22-24 April 2026, i6doc.com publ., ISBN 9782875870964. Available from http://www.i6doc.com/en/.

bootstrap size in terms of out-of-bag estimate of the generalization error is generally lower than the size of the original training set, i.e., $\sharp\mathcal{D}_j < \sharp\mathcal{D}$ [6]. This is because smaller samples tend to increase the diversity of base learners.

Enhancing diversity can also be achieved by forcing the sets $\mathcal{D}_j$ to be disjoint. For instance, empirical evidence suggests that a committee of $J > 1$ base learners fitted on random partitions of $\mathcal{D}$ can outperform a committee created using the same number and size of bootstrap aggregates [7]. Similarly, non-random partitioning methods, where the sets $\mathcal{D}_j$'s are obtained via clustering, also favor diversity. This idea has been investigated in the specific context of Random Forests in regression tasks [8]. The authors fix a number J of (disjoint) clusters $\mathcal{D}_j$, fit a forest on each of them and compute stacked regression weights to combine the forests φ_j's. A different approach is considered in [9], where the authors propose to compute the output $\hat{y}$ associated with a test point x by considering multiple clustering tasks. First, they set a number K of layers. For each layer $k = 1, \ldots, K$, the training set $\mathcal{D}$ is partitioned into k clusters, $C_1^k, \ldots, C_k^k$, and a model φ is fitted on each of them, leading to k base learners $\varphi_1^k, \ldots, \varphi_k^k$. Next, for each layer, they identify the model $\varphi_{i(x)}^k$ associated with the cluster $C_{i(x)}^k$ which the test point x belongs to. Finally, the resulting prediction is given by the simple average of the local predictions:

$$\hat{y} = \sum_{k=1}^{K} \frac{1}{K} \sum_{i=1}^{k} \mathbb{1}_{\{x \in C_i^k\}} \varphi_i^k(x) = \sum_{k=1}^{K} \frac{1}{K} \varphi_{i(x)}^k(x) \ .$$

Note that the first layer ($k = 1$) coincides with the standard fit approach where a single model φ is trained on the set $\mathcal{D}$ available for training. In contrast with [7], however, observe that the clusters $C_{i(x)}^1, \ldots, C_{i(x)}^K$ involved in the combination are related to different layers but to the same test point x, hence, are typically not disjoint. The authors provide theoretical insights about why clustering could be a useful step prior to prediction. The reason for averaging over K layers is to avoid relying on an arbitrary number k of clusters, thereby reducing the sensitivity of the results with respect to K.

In this work, we propose a general framework for cluster aggregating (in short *clagging*) which is a compromise between these clustering-based approaches: It adopts the multi-layer approach proposed in [9] but, as in [8], all the $J = 1 + 2 + \ldots + K = K(K+1)/2$ models intervene when predicting the response of any test point x. In addition, in line with the philosophy of clustering, we replace the binary weights $\mathbb{1}_{\{x \in C_i^k\}}/K$ considered in [9] by continuous variables whose values are related to the distance d between x and each of the J clusters:

$$\hat{y} = \sum_{k=1}^{K} \sum_{i=1}^{k} w_i^k \varphi_i^k(x) \ , \quad w_i^k \propto d(x, C_i^k) \ .$$

Finally, we adapt our framework to accommodate categorical data.

We analyze the performance of our approach via an extensive horse race study on 10 regression datasets with 7 prediction algorithms. We use as benchmarks

ESANN 2026 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 22-24 April 2026, i6doc.com publ., ISBN 9782875870964. Available from http://www.i6doc.com/en/.

not only a standard model fitted on the whole training set (which corresponds to the special case $K = 1$) but also a bagging model that contains a similar number of base learners. Our evidence suggests that relying on clustering to build non-random partitions of the training set boosts the performance when compared to the standard bagging procedure. Moreover, we find that clagging can improve the performance of any prediction algorithm compared to a standard fit on the whole training set in the sense that clagging with $K > 1$ leads to better results compared to choosing $K = 1$, in general.

2 Methodology

The first step of clagging consists in setting the number K of layers and, for each layer $k = 1, \ldots, K$, to apply a clustering algorithm splitting the training set $\mathcal{D}$ in k clusters. By doing so, one obtains K partitions of $\mathcal{D}$ made of an increasing number of sets $\{C_1^k, \ldots, C_k^k\}$, from $k = 1$ to $k = K$. We denote by $\mathcal{S}$ the resulting list of the $J := K(K+1)/2$ subsets of $\mathcal{D}$. In this paper, we focus on k-means, which has become one of the most popular clustering algorithms due to its computational speed, convergence, ease of interpretation and scalability. Moreover, it can create clusters of different shapes and sizes. We rely on K-prototypes when confronted to datasets including both numerical and categorical variables. However, the choice of the clustering algorithm does not seem to have a significant impact on the results [8]. The second step consists in fitting a prediction algorithm φ_i^k on each subset $C_i^k \in \mathcal{S}$. By doing so, we have for each test point x a set of predictions $(\hat{y}_1, ..., \hat{y}_J)^T = \hat{\boldsymbol{y}}$ that we need to combine. The third step consists in aggregating the predictions $\hat{y}_1, \ldots, \hat{y}_J$ computed above into a single response, $\hat{y}$. We restrict ourselves to discussing linear combinations, $\hat{y} = \boldsymbol{w}^T \hat{\boldsymbol{y}}$ where $\boldsymbol{w} \in \mathbb{R}^J$, although non-linear schemes can be considered, too.

For clarity, we order the J models according to the index mapping $(i, k) \to j(i, k) := k(k-1)/2 + i$ for $k = 1, \ldots, K$ and $i = 1, \ldots, k$, such that $\varphi_i^k(x) = \hat{y}_{j(i,k)}$. A first widely used method to determine the weights is the equally-weighted strategy where $w_j^{EW} := 1/J$. Another possibility is to adopt the minimum-variance strategy, aiming to minimize the variance of the error. For instance, it is shown in [10] that minimizing $\mathbb{E} \mid\mid y - \boldsymbol{w}^T \hat{\boldsymbol{y}} \mid\mid_2^2$ under the constraint $\boldsymbol{w}^T \mathbf{1} = 1$ is equivalent to setting:

$$\boldsymbol{w}^{MV} := \underset{\boldsymbol{w}}{\operatorname{argmin}}\ \boldsymbol{w}^T \hat{\boldsymbol{\Sigma}} \boldsymbol{w},$$

where $\hat{\boldsymbol{\Sigma}}$ is the estimated covariance matrix of models' prediction error $v_j := y - \hat{y}_j$. Choosing a robust estimator of the covariance matrix can outperform the robust equally-weighted scheme, as explained in [3]. In the specific setup of clagging, one could adopt the setup proposed in [9] and combine only the predictions associated with models trained on the K clusters (one per layer) which the test point belongs to $w_{j(i,k)}^{TPH} := \mathbb{1}_{\{x \in C_i^k\}}/K$. For each layer k, a test point x belongs to one and only one of those k clusters, $C_{i(x)}^k$, the one for which the

ESANN 2026 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 22-24 April 2026, i6doc.com publ., ISBN 9782875870964. Available from http://www.i6doc.com/en/.

centroid is the closest in terms of distance, $i(x) \coloneqq \text{argmin}_{i\in\{1,\dots,K\}}\, d(x, C_i^k)$. By construction, there will be exactly K ones and $J-K$ zeros in the vector $\boldsymbol{w}^{TPH}$. We also consider setting the weights w_j proportional to the distance between the test point x and the centroids of the corresponding cluster. Denoting by p the number of explanatory variables in the dataset, the aggregated prediction associated with a test point x is given by the combination scheme defined as:

$$w^d_{j(i,k)} \coloneqq \frac{\frac{1}{d(x,C_i^k)}^{p-1}}{\sum_{k=1}^{K}\sum_{i=1}^{k}\frac{1}{d(x,C_i^k))}^{p-1}}.$$

This ensures that the weight vector is defined on the unit simplex.

3 Empirical Analysis

We restrict ourselves to regression due to space constraints, but clagging could be easily extended to classification. We consider 10 different regression datasets coming from UCI repository displayed in Table 1 and 7 different prediction algorithms: Decision Tree (CART), Partial Least Squares, Boosted Generalized Linear Model, Extreme Gradient Boosting, Multivariate Adaptive Regression Spline, Linear Regression and Least Angle Regression. Algorithms are fitted using the `R` package `Caret`. To avoid cherry-picking, we fix the hyperparameters of each algorithm for all the datasets *a priori* using default values from the `R` package `Caret`. For each of the 10 datasets and using the same random seed, we allocate 75% of all observations to a training set to train base learners, 10% of all observations to a validation set to compute the weights and 15% of all observations to a testing set to test the performance of each strategy. Note that only the minimum-variance combination scheme needs a validation set to compute the weights. For the other combination schemes, we merge the training and validation sets.

Dataset	# variables	# observations	Type of variables
Communities&Crime	100	1,994	Numerical
Parkinsons	17	5,875	Numerical
Extreme Weather	22	7,752	Numerical
Electrical Grid	13	10,000	Numerical
Appliances Energy	28	19,735	Numerical
Superconductivity	82	21,263	Numerical
Protein	10	45,730	Numerical
Abalone	9	4,177	Mixed
Seoul	12	8,465	Mixed
Bike Sharing	13	17,379	Mixed

Table 1: Description of the 10 UCI datasets used. *Mixed* type of variables refers to a dataset containing both numerical and categorical features.

ESANN 2026 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 22-24 April 2026, i6doc.com publ., ISBN 9782875870964. Available from http://www.i6doc.com/en/.

First, we test clagging with different combination schemes and we assess if the information that a point might belong to a cluster should be regarded. We compare the performance of a distance-based scheme against the equally-weighted strategy, the minimum-variance strategy and the one proposed in [9]. We consider a sample covariance matrix estimator with a non-negative constraint on the weights for the minimum-variance strategy. We aim to have an order of magnitude for J around 100. This choice follows common practice in ensembling [4, 11]. Therefore we set $K = 14$. For each of the 10 datasets and for each of the 7 prediction algorithms, we fit a model on each of the $J = 105$ subsamples of the training set $\mathcal{D}$. Then, we compare the performance of clagging compared to bagging. For bagging, we consider the most widely used setup in the literature: We create bootstrapped samples by selecting examples from the original training dataset with replacement so that the size of the bootstrapped samples coincides with the size of the original training set. We consider two combination schemes for bagging: equally-weighted strategy and minimum-variance strategy. For a fair comparison, we consider a similar number of samples (100) for bagging. Finally, we compare the performance of clagging compared to a standard fit on the whole training set $\mathcal{D}$.

We use the MSE as performance metric and assess significance using a paired t-test. Table 2 suggests that distance-based combination scheme with clagging outperforms clagging using every other combination scheme. Moreover, distance-based combination scheme with clagging outperforms bagging regardless of the combination scheme considered. For example, comparing clagging with distance-based weights to the approach $\mathcal{M}$ being bagging with equally-weighted scheme, the former performs better than the latter in 51 configurations, worse in 4 configurations, and the two approaches are statistically equivalent in 15 configurations. Finally, distance-based combination scheme with clagging considering a base learner φ performs better or similar to a standard fit of φ on the whole training set except in 2 out of 70 configurations.

$\mathcal{M}$	Clagging(w^d) $>$ $\mathcal{M}$	Clagging(w^d) $\approx$ $\mathcal{M}$	Clagging(w^d) $<$ $\mathcal{M}$
Clagging(w^{EW})	65	4	1
Clagging(w^{MV})	63	6	1
Clagging(w^{TPH})	27	37	6
Bagging(w^{EW})	51	15	4
Bagging(w^{MV})	53	15	2
Standard fit	54	14	2

Table 2: Column 1 reports the number of configurations where clagging with distance-based combination scheme is significantly better than clagging with other schemes. Column 2 reports the number of configurations where there is no significant difference between strategies. Column 3 reports the number of configurations where distance-based combination scheme is significantly worse. Performance metric considered is Mean Squared Error with a paired t-test at a 95% confidence level to assess significance.

ESANN 2026 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 22-24 April 2026, i6doc.com publ., ISBN 9782875870964. Available from http://www.i6doc.com/en/.

4 Conclusion

Ensemble learning has been shown to outperform individual models, especially when there is some diversity in the set of base learners. One of the most widely used ensemble learning model is *Bagging*. Usually, a bootstrapped sample is created by drawing samples from the original training dataset with replacement so that the size of the bootstrapped samples coincides with the size of the original training set. In this work, we propose to aggregate clusters instead of bootstraps. The general idea of cluster aggregating (clagging) is to rely on clustering to divide the training set into clusters on which prediction models are fitted. Predictions of the different models are subsequently combined and the information that a point might belong to a cluster should not be disregarded. A combination scheme based on the distance of a test point to each cluster is likely to outperform standard benchmarks such as equally-weighted and minimum-variance strategies. Our evidence obtained through a horse race study with 10 datasets and 7 prediction algorithms suggests that clagging outperforms the standard random sampling with replacement widely used for bagging. Moreover, clagging outperforms a standard fit on the whole training set.

References

[1] Ludmila I Kuncheva and Christopher J Whitaker. Measures of diversity in classifier ensembles and their relationship with the ensemble accuracy. *Machine learning*, 51:181–207, 2003.

[2] Amir F Atiya. Why does forecast combination work so well? *International Journal of Forecasting*, 36(1):197–200, 2020.

[3] Francesco Roccazzella, Paolo Gambetti, and Frédéric Vrins. Optimal and robust combination of forecasts via constrained optimization and shrinkage. *International Journal of Forecasting*, 38(1):97–116, 2022.

[4] Leo Breiman. Bagging predictors. *Machine learning*, 24:123–140, 1996.

[5] Peter Bühlmann and Bin Yu. Analyzing bagging. *The annals of Statistics*, 30(4):927–961, 2002.

[6] Gonzalo Martínez-Muñoz and Alberto Suárez. Out-of-bag estimation of the optimal sample size in bagging. *Pattern Recognition*, 43(1):143–152, 2010.

[7] Nitesh Chawla, Thomas E Moore, Kevin W Bowyer, Lawrence O Hall, Clayton Springer, and Philip Kegelmeyer. Bagging is a small-data-set phenomenon. In *Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. CVPR 2001*, volume 2, pages II–II. IEEE, 2001.

[8] Maya Ramchandran, Rajarshi Mukherjee, and Giovanni Parmigiani. Cross-cluster weighted forests. *arXiv preprint arXiv:2105.07610*, 2021.

[9] Shubhendu Trivedi, Zachary A Pardos, and Neil T Heffernan. The utility of clustering in prediction tasks. *arXiv preprint arXiv:1509.06163*, 2015.

[10] Clive WJ Granger and Ramu Ramanathan. Improved methods of combining forecasts. *Journal of forecasting*, 3(2):197–204, 1984.

[11] Trevor Hastie, Robert Tibshirani, Jerome Friedman, et al. The elements of statistical learning, 2009.