Generating experts by boosting diversity

Jean-Michel Poggi[1], Mathias Bourel[2], Jairo Cugliari[3], Yannig Goude[4]

[1] LMO, Univ. Paris-Saclay, Orsay & Univ. Paris, France – Jean-Michel.Poggi@math.u-psud.fr
[2] Universidad de la Republica, Montevideo, Uruguay– mbourel@fing.edu.uy
[3] University Lumière Lyon 2, France – Jairo.Cugliari@univ-lyon2.fr
[4] EDF, Paris-Saclay, France – yannig.goude@edf.fr

Abstract

The practical interest of using ensemble methods has been highlighted in several works. Aggregation estimation as well as sequential prediction provide natural frameworks for studying ensemble methods and for adapting such strategies to time series data. Sequential prediction focuses on how to combine by weighting a given set of individual experts while aggregation is mainly interested in how to generate individual experts to improve prediction performance.

We look for enhancing these (possibly online) mixture methods by using the concept of diversity. We propose an algorithm to enrich the set of original individual predictors using a gradient boosting-based method by incorporating a diversity term to guide the gradient boosting iterations. The idea is to progressively generate experts by boosting diversity.

We establish a convergence result ensuring that the associated optimization strategy converges to a global optimum. Then, we show by means of numerical experiments the appropriateness of our procedure using simulated data and real-world electricity demand datasets.

Keywords: Diversity, Boosting, Forecasting.

1 Introduction

The practical interest of using ensemble methods has been highlighted in several works (see [12]). These studies focus on the rules of aggregation of a set of experts and examine how to weight and combine the experts. Ensemble methods are now used in very different domains (see [11]).

Boosting techniques are iterative methods that consist in improving the performance of several hypothesis or base predictors of the same nature, combining them and re-weighting at each step the original data sample. Freund and Schapire in [6] described the first boosting algorithm, Adaboost, designed for binary classification problems and with classification trees as hypothesis. Various types of extensions for boosting exist, in particular for multi-class classification and for regression and they use different approaches [14].

We use the concept of diversity [4, 13] to propose in the regression context, an algorithm to enrich the set of original individual predictors. This formulation is inspired from the Negative Correlation Learning for neural networks [9]. We modify the usual L^2 cost function with the aim to find a good predictor that will be at the same time "diverse" than the mean of the predictors founded at the precedent steps, according to the diversity formula.

We establish a convergence result ensuring that the associated optimization strategy converges to a global optimum. Then, we show by means of numerical experiments the appropriateness of our procedure using simulated data and real-world electricity demand datasets.

2 Diversity decomposition of a set of experts

The decomposition below is true for any convex aggregation rules but we will take here uniform weights (since it is the usual output of boosting algorithm iterations). Consider a prediction $\hat{y}_i$ which is the simplest convex aggregation of a set of M individual experts $\hat{y}_{i,m}$: $\hat{y}_i = \frac{1}{M}\sum_{m=1}^{M}\hat{y}_{i,m}$. This special kind of mixture gives particularly nice expressions when one decomposes the instantaneous square error $(y_i - \widehat{y}_i)^2$, as proposed in [4], with the diversity formula:

$$(y_i - \hat{y}_i)^2 = \underbrace{\frac{1}{M}\sum_{m=1}^{M}(\hat{y}_{i,m} - y_i)^2}_{\text{weighted average error of the individuals}} - \underbrace{\frac{1}{M}\sum_{m=1}^{M}(\hat{y}_{i,m} - \widehat{y}_i)^2}_{\text{diversity term}} \tag{1}$$

The significance of the Ambiguity decomposition it that the error of the ensemble will be less than or equal to the average error of the individuals, and then the ensemble has lower error than the average individual error: larger will be the diversity term, larger will be the ensemble error reduction.

3 Diversity-based cost function

In the context of machine learning methods, boosting are sequential algorithms that estimate a function $F : \mathbb{R} \to \mathbb{R}$ by minimizing the expectation of a functional $C(F) = \mathbb{E}\left[\Psi(Y, F(X)\right]$ where $\Psi : \mathbb{R} \times \mathbb{R} \to [0, +\infty)$ measures the cost for predicting $F(X)$ instead of Y, using a training sample $\{(y_i, x_i)\}_{i=1}^{n}$ and functional gradient descent techniques. More precisely, considering a family $\mathcal{F} = \{f : \mathbb{R}^d \to \mathbb{R}\}$, the method consists to estimate F by minimisation of the empirical expectation loss

$$C_n(F) = \frac{1}{n}\sum_{i=1}^{n}\Psi(y_i, F(x_i)), \tag{2}$$

by looking for an additive function of the form $F_M = \sum\limits_{m=1}^{M}\alpha_m f_m$ where $\alpha_m \in \mathbb{R}$ and $f_m \in \mathcal{F}$ for all m (see [8], [10] and [5]). The general Gradient Boosting algorithm is recalled in Figure 1.

In the spirit of L^2-Boost algorithm, we propose an algorithm which encourage diversity of intermediate predictors. Following equation (1), we propose the cost function

$$\Psi(y_i, F) = \frac{1}{2}(y_i - F)^2 - \frac{\kappa}{2}(F - c)^2$$

where κ is the term which modulate the importance given to the diversity of the predictor to the average of the previous ones and therefore c can be thought as a constant. The **Bo**osting **Di**versity algorithm (Bodi) is detailed in Figure 2.

With BoDi algorithm, as in the classical boosting, we obtain a final ensemble forecast $F^*_{M,\kappa}$ as well as a set of experts $F_{k,\kappa}$. We make the dependency to κ explicit whereas other parameters (like the gradient step, the size of the bootstrap sample) play a role.

4 A convergence result

An elegant and recent result from [1] can be used to prove the convergence of several gradient boosting-based methods in a very general framework. The convergence result holds for $C(F)$, the

```
Let $\mathcal{L} = \big\{(y_1, \mathbf{x}_1), \dots, (y_n, \mathbf{x}_n)\big\}$ be a sample and $\mathcal{F}$ a family of functions.

  1. Fit an initial learner $\widehat{f}_0 = \widehat{F}_0 \in \mathcal{F}$ such that $\widehat{F}_0 = \operatorname*{Argmin}_{f \in \mathcal{F}} \sum_{i=1}^{n} \big(y_i - f(\mathbf{x}_i)\big)^2$

  2. For $m \in \{1, \dots, M\}$:

      (a) Compute the negative gradient $u_i = -\frac{\partial \Psi(y_i,F)}{\partial F}\Big|_{F=\widehat{F}_{m-1}(\mathbf{x}_i)}$ $\forall\, i = 1, \dots, n$ and
          $\widehat{f}_m = \operatorname*{Argmin}_{f \in \mathcal{F}} \sum_{i=1}^{n} \big(u_i - f(\mathbf{x}_i)\big)^2$

      (b) Choose the best step size: $\widehat{w}_m = \operatorname*{Argmin}_{w} \sum_{i=1}^{n} \Psi\big(y_i, \widehat{F}_{m-1}(\mathbf{x}_i) + w\widehat{f}_m(\mathbf{x}_i)\big)$

      (c) Update the aggregated predictor: $\widehat{F}_m(x) = \widehat{F}_{m-1}(\mathbf{x}) + \widehat{w}_m \widehat{f}_m(\mathbf{x})$

Outputs: a family of experts $\widehat{f}_0, \widehat{f}_1, \widehat{f}_2, \dots, \widehat{f}_M$ and the aggregated predictor $\widehat{F}_M$.
```

Figure 1: General Gradient Boosting algorithm.

```
Let $\mathcal{L} = \big\{(y_1, \mathbf{x}_1), \dots, (y_n, \mathbf{x}_n)\big\}$ be a sample, $\mathcal{F}$ a family of functions, $\kappa > 0$ and $\delta > 0$. Randomly
split the data in two disjoint parts $I = I_1 \cup I_2$.

  1. Fit an initial learner $\widehat{f}_0 = \widehat{F}_0 \in \mathcal{F}$ over $I_1$ such that $\widehat{F}_0 = \operatorname*{Argmin}_{f \in \mathcal{F}} \sum_{i \in I_1} \big(y_i - f(\mathbf{x}_i)\big)^2$.
     Set $\widehat{F}_0^*(\mathbf{x}) = \widehat{F}_0(\mathbf{x})$.

  2. For $m \in \{1, \dots, M\}$:

      (a) $\forall\, i \in I_2$, compute the negative diversity gradient of the cost evaluated at $\widehat{F}_{m-1}(\mathbf{x}_i)$:

              $u_i = \big(y_i - \widehat{F}_{m-1}(\mathbf{x}_i)\big) + \kappa_m \Big(\widehat{F}_{m-1}(\mathbf{x}_i) - \widehat{F}_{m-1}^*(\mathbf{x}_i)\Big)$

          with $\kappa_m = \kappa\left(1 - \frac{1}{m}\right)$ if $m > 1$, $\kappa_1 = \kappa$ and $\widehat{f}_m = \operatorname*{Argmin}_{f \in \mathcal{F}} \sum_{i \in I_2} \big(u_i - f(\mathbf{x}_i)\big)^2$.

      (b) Update boosting predictor as $\widehat{F}_m(\mathbf{x}) = \widehat{F}_{m-1}(\mathbf{x}) + \delta \widehat{f}_m(\mathbf{x})$.
          Compute $\widehat{F}_m^*(\mathbf{x}) = \frac{1}{m} \sum_{i=1}^{m} \widehat{F}_i(\mathbf{x})$ and update $I_2 = I \setminus I_1$ with a new subsample $I_1$ of $I$.

Outputs: a family of experts $\widehat{f}_0, \widehat{f}_1, \widehat{f}_2, \dots, \widehat{f}_M$ and the aggregated predictors $\widehat{F}_M$ and $\widehat{F}_M^*$.
```

Figure 2: Boosting Diversity algorithm.

expectation of our convex cost function $\Psi(y, F) = \frac{1}{2}(y - F)^2 - \frac{\kappa}{2}(F - c)^2$ because C and Ψ satisfies the three assumptions needed to ensure convergence established. It warranties that the optimisation strategy converges to a global optimum. It should be noted that this result is not a statistical one.

5 Numerical experiments

A simulated example We use here a simulated data set presented in [7] and used in [3] for bagging. The inputs are 10 independent variables uniformly distributed on the interval [0,1], only 5 out of these 10 are actually used. Outputs y are generated according to the formula:

$$y_i = 10\sin(\pi x_{1,i}x_{2,i}) + 20(x_3 - 0.5)^2 + 10x_4 + 5x_5 + \varepsilon_i$$

where ε_i is $N(0, \sigma^2)$. As in [3] we simulated a learning set of size $n_0 = 200$ and a test set of size $n_1 = 1000$ observations, $\sigma = 1$. We replicate the simulation 100 times. The results for a random forest base learner including all the 10 covariates with parameters $\texttt{mtry} = 3$, $\texttt{ntree} = 100$, data splitting rate $\alpha = 0.5$, gradient step $\delta = 0.08$ and diversity weight $\kappa \in \{0, 0.5, 1, 1.5\}$ are presented in Figure 3.

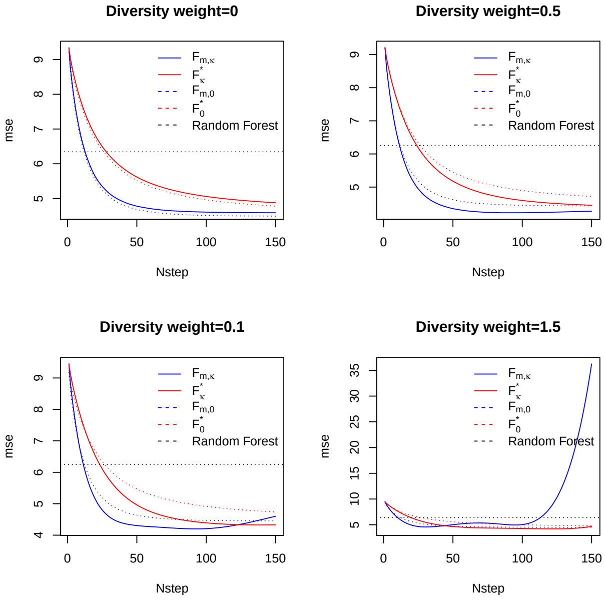

Figure 3: MSE for different diversity (κ) values with RF ($\texttt{mtry}$=3, $\texttt{ntree}$=100) as base learner.

We see the influence of κ. For κ not too large there is a clear improvement of the diversity boosting strategy over the original random forest forecaster, reducing the error by 3 after a sufficient number of iterations (at least 100). For large κ, here 1.5 or more, the algorithm diverges after 100 iterations. In the range of reasonable values of κ (ensuring convergence of the algorithm), choosing κ too small entails a larger forecasting error meaning that encouraging diversity can lead to an improvement of the forecasts. $\kappa = 0$ corresponds to classical boosting. We can observe that classical boosting works well here and improves significantly the forecast of the original forest. This is quite surprising since the random forest could be seen as a ”strong” learner in the sense that it is not a weak learner as stumps or small trees or other classical weak learners in boosting.

Three base-learners are considered : stumps, Purely Random Forests (PRF) and Breiman's Random Forests (RF). To illustrate the influence of the choice of the base learner, let us examine the MSE as a function of boosting steps for the 3 base learners (see Figure 4).

The best results are obtained for RF. But the relative improvement over the original base learner is far more important for PRF, probably because PRF can generate more diversity than RF, inducing

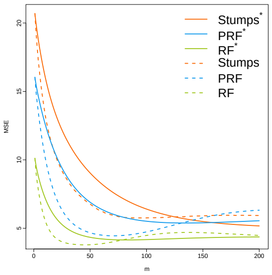

Figure 4: MSE as a function of boosting steps for the 3 base learners: Stumps, PRF and RF.

a large gain with diversity boosting. Another remark is the good convergence of the algorithm and its robustness regarding the choice of the number of boosting steps.

An electricity consumption dataset Consider French electricity consumption provided by the system operator RTE (Reseau de Transport d'Electricite), from the 1st of January 2012 to the 15th of March 2020 with a 30 minutes sampling period. We add a covariate: thenational averaged temperature from the French weather forecaster Meteo-France. We train the models on historical data from January 2012 to the end of August 2019 and test on the last year. Finally, to avoid outliers we drop the holidays periods and bank holidays.
Examine first, the performance by considering RF as base learner in Figure 5.

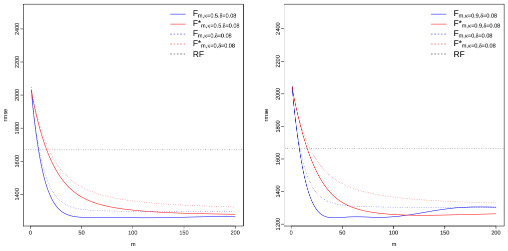

Figure 5: RMSE on test set as a function of boosting steps for $\kappa = 0.5$ (left) $\kappa = 0.9$ (right) gradient steps ($\delta = 0.08$) with RF (ntree=100, mtry=3) as base learner.

The best RMSE are obtained for F_κ, followed by F^*_κ. Choosing κ close to 1 leads to improve forecasting performance as the learner can generate more diversity.
To end, examine the performance by considering PRF as base learner in Figure 6. Interestingly, the best RMSE is achieved by $F_{\kappa=0.9}$ with PRF which is a good base learner for diversity boosting.

Acknowledgements This work benefited from the support of the PGMO / IRSDI program (`https://www.fondation-hadamard.fr/en`).

Notice This text is an extended abstract of a full paper submitted for publication (see [2]).

References

[1] G. Biau and B. Cadre. Optimization by gradient boosting. *Preprint arXiv:1707.05023*, 2017.

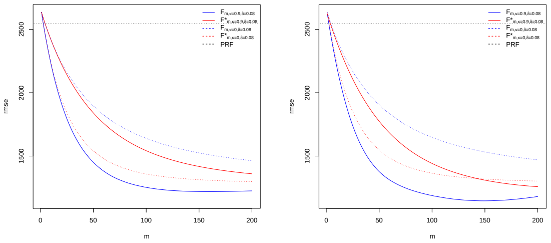

Figure 6: RMSE on test set as a function of boosting steps for $\kappa = 0.5$ (left) $\kappa = 0.9$ (right) gradient steps ($\delta = 0.08$) with PRF (`ntree`=100) as base learner.

[2] M. Bourel, J. Cugliari, Y. Goude, and J.-M. Poggi. Boosting diversity in regression ensembles. *Preprint, https://hal.archives-ouvertes.fr/hal-03041309*, 2020.

[3] L. Breiman. Bagging predictors. *Machine learning*, 24(2):123–140, 1996.

[4] G. Brown, J. L. Wyatt, and P. Tiňo. Managing diversity in regression ensembles. *J. Mach. Learn. Res.*, 6:1621–1650, Dec. 2005.

[5] P. Bühlmann and B. Yu. Boosting with the l2 loss. *Journal of the American Statistical Association*, 98(462):324–339, 2003.

[6] Y. Freund and R. Schapire. A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of Computer and System Sciences*, 55(1):119–139, 1997.

[7] J. Friedman. Multivariate adaptive regression splines. *The annals of statistics*, 19(1):1–67, 1991.

[8] J. Friedman, T. Hastie, and R. Tibshirani. Additive logistic regression: a statistical view of boosting (with discussion and a rejoinder by the authors). *The Annals of Statistics*, 28(2):337–407, 2000.

[9] Y. Liu and X. Yao. Ensemble learning via negative correlation. *Neural Netw.*, 12(10):1399–1404, Dec. 1999.

[10] L. Mason, J. Baxter, P. Bartlett, and M. Frean. Functional Gradient Techniques for Combining Hypotheses. In *Advances in Large-Margin Classifiers*. The MIT Press, 09 2000.

[11] R. Polikar. Ensemble based systems in decision making. *IEEE Circuits and systems magazine*, 6(3):21–45, 2006.

[12] R. Polikar. Ensemble learning. In *Ensemble machine learning*, pages 1–34. Springer, 2012.

[13] H. W. Reeve and G. Brown. Diversity and degrees of freedom in regression ensembles. *Neurocomputing*, 298:55 – 68, 2018.

[14] R. Schapire and Y. Freund. *Boosting: Foundations and Algorithms*. Adaptive Computation and Machine Learning Series. Mit Press, 2012.