Generating experts by boosting diversity

Mathias Bourel, Jairo Cugliari, Yannig Goude, Jean-Michel Poggi

Generating experts by boosting diversity

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

[1] *Universidad de la República, Uruguay,* mbourel@fing.edu.uy
[2] *Université Lumière Lyon 2, France,* jairo.cugliari@univ-lyon2.fr
[3] *EDF R&D, Université Paris Sud, France,* yannig.goude@edf.fr
[4] *Université Paris Descartes, Université Paris Sud, France,*
jean-michel.poggi@math.u-psud.fr

Résumé. L'intérêt pratique de l'utilisation de méthodes d'ensemble a été souligné dans plusieurs ouvrages. La prévision séquentielle de suites individuelles fournit un cadre naturel pour adapter les méthodes d'ensemble aux séries chronologiques. Les principaux développements dans ce domaine concernent les algorithmes d'agrégation et la mise à jours des poids en ligne. Nous nous concentrons ici sur la question de la génération des experts à inclure dans une agrégation en ligne. Nous exploitons pour cela le concept de diversité pour proposer des stratégies d'enrichissement d'un ensemble d'experts initiaux. Nous montrons comment cette approche s'incrit dans les travaux théoriques récents sur le boosting. Nous proposons des résultats de simulations montrant la pertinance de cette approche en régression. Des applications sur des données réelles (consommation électrique et pollution) confirment l'intérêt de cette méthodes pour la prévision de séries temporelles.

Mots-clés. agrégation d'experts, boosting, prévision.

Abstract. The practical interest of using ensemble methods has been highlighted in several works. Sequential prediction provides a natural framework for adapting ensemble methods to time series data. Main developments focus on the rules of aggregation of a set of experts and examine how to weight and combine the experts. However, very few work exist regarding how to choose/generate the experts to include in a given aggregation procedure. We use the concept of diversity to propose some strategies to enrich the set of original individual predictors. We show how this method is connected to recent theoretical work on boosting. We propose a simulation study to illustrate the interest of our approach in the regression setting. An application on real datasets (electricity consumption and pollution data) shows the potentiality of this method for practical forecasting tasks.

Keywords. expert aggregation, boosting, forecasting.

1 Introduction

The practical interest of using ensemble methods has been highlighted in several works. 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. [1] provided with global comparison of different approaches on ecology. 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 [5] 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 ([12]). We use the concept of diversity [3, 11] to propose new algorithms to enrich the set of original individual predictors. This formula is inspired from the Negative Correlation Learning for neural networks ([9]). 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. 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 show by means of numerical experiments the appropriateness of our procedure using simulated data and electricity demand datasets.

2 Diversity decomposition of a set of experts

We consider here a prediction $\hat{y}_i$ which is the 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 [3], with the diversity formula:

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

This decomposition is true for any convex aggregation rules but for simplicity we will consider here uniform weights.

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 minimising 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 committed 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 the empirical expectation loss

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

by looking for an additive function of the form $F_M = \sum_{m=1}^{M} \alpha_m f_m$ where $\alpha_m \in \mathbb{R}$ and $f_m \in \mathcal{F}$ for all m ([6], [10],[4]).

In the spirit of L^2-Boost algorithm, we propose a new algorithm which encourage diversity of intermediate predictors. Following equation (1), we propose as new 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 think as a constant. Let $\mathcal{L} = \{(y_1, x_1), \ldots, (y_n, x_n)\}$ be a sample with unknown distribution F and consider a fix step $\delta > 0$. Then the **Bo**osting **Di**versity algorithm (Bodi) is detailed in Figure 1.

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. This is because we want to study more in depth the interest of the diversity term for forecasting purposes.

4 Results

Convergence of the algorithm. A recent and very elegant result from [4, Theo. 1] proves the convergence of several gradient boosting-based methods in a very general framework. The results holds for $C(F)$, the 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. This convergence result warranties that the optimisation strategy converges to a global optimum.

Remark: usual boosting methods consider as base-learner F a weak (but computationally efficient) method (e.g. stumps or single variable regression), denoted by h indexed by $\theta \in \Theta$. In our experiments, we show that, to be able to generate some diversity, F has to be a more complex learner. We obtained good result choosing a random forest as base learner.

Given a base learner h, randomly split the data in two parts $I = I_1 \cup I_2$.

1. Fit an initial learner over I_1: $\widehat{F}_0(x) = h(x, \widehat{\theta}_{Y,X})$ where $\widehat{\theta}_{Y,X} = \arg\min_\theta \sum_{i=1}^n (y_i - h(x_i, \theta))^2$. Set $\widehat{F}_0^*(x) = \widehat{F}_0(x)$.
2. For $m \in \{1, \ldots, M\}$:
   (a) $\forall\, i \in I_2$, compute the negative diversity gradient of the cost function and evaluate it at $\widehat{F}_{m-1}(x_i)$:
   $$u_i = (y_i - \widehat{F}_{m-1}(x_i)) + \kappa(\widehat{F}_{m-1}(x_i) - \widehat{F}^*_{m-1}(x_i))$$
   and compute $\widehat{g}_m(x) = h(x, \widehat{\theta}_{U,X})$ where $\widehat{\theta}_{U,X} = \arg\min_\theta \sum_{i=1}^n (u_i - h(x_i, \theta))^2$.
   (b) Update boosting predictor as $\widehat{F}_m(x) = \widehat{F}_{m-1}(x) + \delta \widehat{g}_m(x)$, compute $\widehat{F}^*_m(x) = \frac{1}{m} \sum_{i=1}^m \widehat{F}_i(x)$ and update $I_2 = I \setminus I_1$ with a new bootstrap sample I_1 of I.

Figure 1: Boosting Diversity algorithm. The split of the original sample is not mandatory, but it allows to compute the OOB error. To avoid surnotation we do not include the dependence of κ.

Some numerical experiments We use here a simulated data set presented in [7] and used in [2] to demonstrate the good performances of bagging. We use the R package [8] to reproduce these datasets. 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 [2] 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 performances reported by Breiman's bagging predictor in terms of mean square error (MSE) on the test set is 6.2. The first test was conducted with the following inputs: base learner is a random forest including all the 10 covariates with parameters $\mathsf{mtry} = 3$, $\mathsf{ntree} = 100$, data splitting rate $\alpha = 0.5$, gradient step $\delta = 0.08$ and diversity weight $\kappa \in \{0, 0.5, 1, 1.5\}$. The results are presented in Figure 2.

We clearly see the influence of the calibration 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

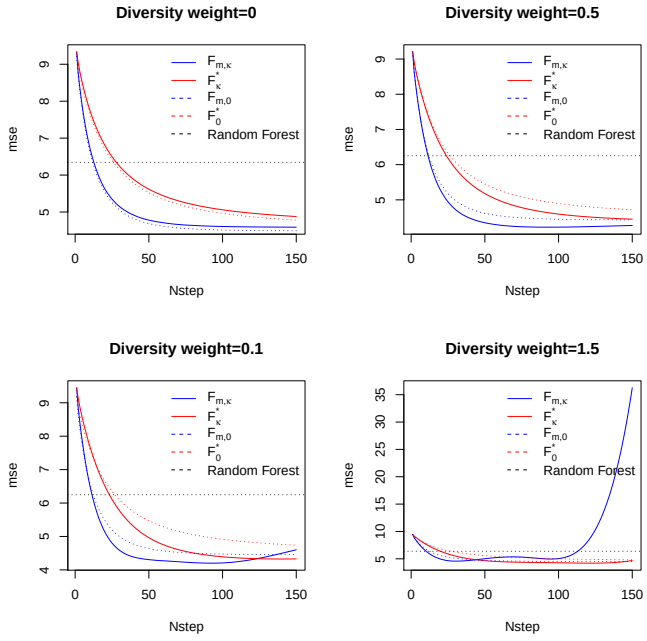

Figure 2: MSE in function of boosting steps for different diversity (κ) values with random forest (mtry=3, ntree=100) as base learner.

κ, here corresponding to 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 also surprising as the random forest could be seen as a "strong" learner in the sense that it is not a weak learner as stumps or other classical weak learners in boosting. Curiously, even if the diversity boosting has been derived so that to improve the ensemble forecast $F^*_{M,\kappa}$ we observe that for all κ and $N < 100$ the error of $F_{M,\kappa}$ is lower than the error of $F^*_{M,\kappa}$.

5 Conclusion

In this work, we propose a new boosting algorithm for regression problems based on the diversity formula. This method constructs at each step a base learner improving the diversity term of the diversity formula and then, try to reduce the mean square error. First experiments on simulated data and random forests as base learner confirm the potentiality of the method. We will now going to apply and adapt it on practical forecasting tasks, examining in particular on how adapting the weights of the combination of the experts.

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

References

[1] M. Bourel, C. Crisci, and A. Martínez. Consensus methods based on machine learning techniques for marine phytoplankton presence–absence prediction. *Ecological Informatics*, 42:46 – 54, 2017.

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

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

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

[5] 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.

[6] 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.

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

[8] F. Leisch and E. Dimitriadou. mlbench—a collection for artificial and realworld machine learning benchmarking problems. r package, version 0.5-6, 2001.

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

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