ESANN 2012 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 25-27 April 2012, i6doc.com publ., ISBN 978-2-87419-049-0.
Available from http://www.i6doc.com/en/livre/?GCOI=28001100967420.

L_1-based compression of random forest models

Arnaud Joly, François Schnitzler, Pierre Geurts and Louis Wehenkel

University of Liège - Department of EE and CS & GIGA-research
Liège, Sart-Tilman, B-28, B-4000, Belgium
{a.joly,fschnitzler,p.geurts,l.wehenkel}@ulg.ac.be

Abstract. Random forests are effective supervised learning methods applicable to large-scale datasets. However, the space complexity of tree ensembles, in terms of their total number of nodes, is often prohibitive, specially in the context of problems with very high-dimensional input spaces. We propose to study their compressibility by applying a L_1-based regularization to the set of indicator functions defined by all their nodes. We show experimentally that preserving or even improving the model accuracy while significantly reducing its space complexity is indeed possible.

1 Introduction

High-dimensional supervised learning problems, *e.g.* in image exploitation and bioinformatics, are more frequent than ever. Tree-based ensemble methods, such as random forests [1] and extremely randomized trees [2], are effective variance reduction techniques offering in this context a good trade-off between accuracy, computational complexity, and interpretability. The number of nodes of a tree ensemble grows as nM (n being the size of the learning sample and M the number of trees in the ensemble). Empirical observations show that the variance of individual trees increases with the dimension p of the original feature space used to represent the inputs of the learning problem. Hence, the number $M(p)$ of ensemble terms yielding near-optimal accuracy, which is proportional to this variance, also increases with p. The net result is that the space complexity of these tree-based ensemble methods will grow as $nM(p)$, which may jeopardize their practicality in large scale problems, or when memory is limited.

While pruning of single tree models is a standard approach, less work has been devoted to pruning ensembles of trees. Reference [3] proposes however to transpose the classical cost-complexity pruning of individual trees to ensembles. On the other hand, references [4, 5, 6] propose to improve model interpretability by selecting optimal rule subsets from tree-ensembles. Another approach to reduce complexity and/or improve accuracy of tree-ensembles is to merely select an optimal subset of trees from a very large ensemble generated in a random fashion at the first hand (see, e.g. [7, 8]).

To further investigate the feasibility of reducing the space complexity of tree-based ensemble models, we consider in this paper the following experiment: (i) build an ensemble of trees; (ii) apply to this ensemble a 'compression step' by reformulating the tree-ensemble based model as a linear model in terms of node indicator functions and by using an L_1-norm regularization approach - à la Lasso [9] - to select a minimal subset of these indicator functions while maintaining predictive accuracy. We propose an algorithmic framework and an empirical investigation of this idea, based on three complementary datasets, and we show that indeed it is possible to so compress significantly tree-based ensemble models, both in regression and in classification problems. We also

ESANN 2012 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 25-27 April 2012, i6doc.com publ., ISBN 978-2-87419-049-0.
Available from http://www.i6doc.com/en/livre/?GCOI=28001100967420.

observe that the compression rate and the accuracy of the compressed models further increase with the ensemble size M, even beyond the number $M(p)$ of terms required to ensure convergence of the variance reduction effect.

The rest of this paper is organized as follows: Section 2 introduces extremely randomized trees and their L_1-norm based compression; Section 3 provides our empirical study and Section 4 concludes and describes further perspectives.

2 Compressing tree ensembles by $L1$-norm regularization

We use the extremely randomized tree algorithm (Extra-Trees, [2]) which builds an ensemble of M trees from a dataset of input-output pairs $((x_i, y_i))_{i=1}^n \in (\mathcal{X} \times \mathcal{Y})^n$ as follows: for each node at which the sample size is greater or equal to a pre-pruning parameter $n_{\min}$, the best split is chosen among a random subset of K variables combined with a random cut point. Setting parameter K to p allows to filter out irrelevant variables, $n_{\min}$ controls the tree complexity possibly at the price of higher bias, and the higher M the smaller the variance.

From an ensemble of M trees, one can extract a set of node indicator functions as follows: each indicator function $1_{m,l}(x)$ is a binary variable equal to 1 if the input vector x reaches the lth node in the mth tree, 0 otherwise. Using these indicator functions, the output predicted by the model may be rewritten as:

$$\hat{y}(x) = \frac{1}{M} \sum_{m=1}^{M} \sum_{l=1}^{N_m} w_{m,l}\, 1_{m,l}(x), \tag{1}$$

where N_m is the number of nodes in the mth tree and $w_{m,l}$ is equal to the leaf-label if node (m, l) is a leaf and to zero if it is an internal node. We can therefore interpret the tree building algorithm as the (random) inference of a new representation which lifts the original input space $\mathcal{X}$ towards $\mathcal{Z}$ of dimension $q = \sum_{m=1}^{M} N_m$ by $z(x) = (1_{1,1}(x), \ldots, 1_{1,N_1}(x), \ldots, 1_{M,1}(x), \ldots, 1_{M,N_M}(x))$.

We propose to compress the tree ensemble by applying a variable selection method to its induced feature space $\mathcal{Z}$. Namely, by $L1$-regularization we can search for a linear model by solving the following optimization problem:

$$\left(\beta_j^*(t)\right)_{j=0}^q = \arg\min_{\beta} \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{q} \beta_j\, z_j(x_i) \right)^2 \text{ s.t. } \sum_{j=1}^{q} |\beta_j| \leq t. \tag{2}$$

This optimization problem, also called LASSO [9], has received much attention in the past decade and is particularly successful in high dimension. The $L1$-norm constraint leads to a sparse solution: only a few weights β_j will be non zero, and their number tends to zero with $t \to 0$; the optimal value t^* of t is problem specific and is typically adjusted by cross-validation. In order to solve Eqn. (2) for growing values of t, we use the 'incremental forward stagewise regression' algorithm [10] which imposes that each $\beta_j^*(t)$ increases monotonically with t. This version deals indeed better with many correlated variables, which is relevant in our setting, since each node indicator function is highly correlated with those of its neighbor nodes in the tree from which it originates. The final weights $\beta_j^*(t^*)$ may be exploited to prune the randomized tree ensemble: a test node can be deleted if all its descendants correspond to $\beta_j^*(t^*) = 0$.

ESANN 2012 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 25-27 April 2012, i6doc.com publ., ISBN 978-2-87419-049-0.
Available from http://www.i6doc.com/en/livre/?GCOI=28001100967420.

3 Empirical analysis

In the following experiments, datasets are pre-whitened: input/output data are translated to zero mean and rescaled to unit variance. All results shown are averaged over 50 experiments in order to avoid randomization artifacts.

When using the Lasso, the incremental forward stagewise algorithm was applied with a 0.01 step size, and the optimal point t^* of the regularization path was chosen by ten-fold cross-validation t^*_{cv} over the learning set (to this end, we used a quadratic loss in regression and a $0-1$ loss in classification).

Below, we will abbreviate extremely randomized trees by "ET" and their $L1$-regularization-based compressed version by "rET".

3.1 Overall performances

We have evaluated our approach on three datasets:

- Friedman1 [11] is a regression problem with $p = 10$ independent input variables of uniform distribution $\mathcal{U}(0,1)$. We try to estimate the output $y = 10\sin(\pi\, x_1\, x_2) + 20(x_3 - \frac{1}{2})^2 + 10x_4 + 6x_5 + \epsilon$, where ϵ is a Gaussian noise $\mathcal{N}(0,1)$. There are 300 learning samples and 2000 testing samples.

- Two-norm [12] is a binary classification problem with $p = 20$ normally distributed (and class-conditionally independent) input variables: either from $\mathcal{N}(-a,1)$ if the class is 0 or from $\mathcal{N}(a,1)$ if the class is 1 (with $a = \frac{2}{\sqrt{20}}$). There are 300 learning and 2000 testing samples.

- SEFTi [13] is a (simulated) regression problem which concerns the tool level fault isolation in a semiconductor manufacturing. One quarter of the values are missing at random and were replaced by the median. There are $p = 600$ input variables, 2000 learning samples and 2000 testing samples.

We have used a set of representative meta-parameter values (K, $n_{\min}$ and M) of the Extra-Trees algorithm (see Table 1). Accuracies are measured on the test sample and complexity is measured by the number of test nodes of the ET and rET models (the compression factor being the ratio of the former to the latter). We observe a compression factor between 9 and 34, a slightly lower error for the rET model than for the ET model on the two regression problems (Friedman1 and SEFTi) and the opposite on Two-norm. To compare, we show the results obtained with the Lasso on the original features: it is much less accurate than both ET and rET on the (non-linear) regression problems (Friedman1 and SEFTi), but superior on the (linear) classification problem (Two-norm).

Side experiments (results not provided) show that changing the value of parameter K does not influence significantly the final accuracy and complexity on the Two-norm and Friedman1 datasets, while for SEFTi, accuracy increases strongly with K (presumably due to a large number of noisy and/or irrelevant features) with however little impact on the final complexity.

3.2 Detailed analysis of the compression method behavior

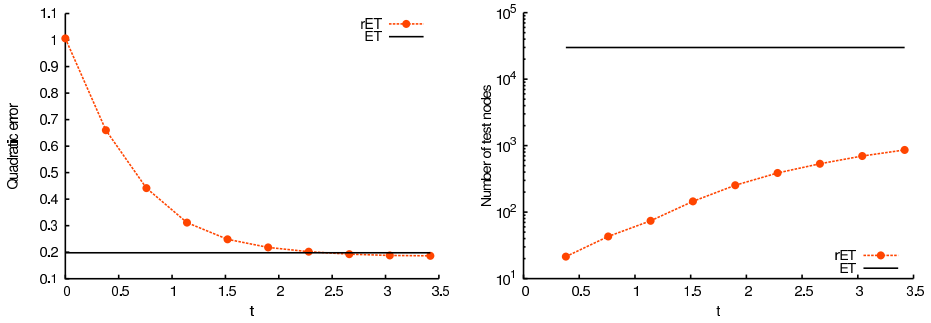
In this section, we further analyze the models obtained on the Friedman1 problem. Similar conclusions can also be drawn for Two-norm and SEFTi datasets.

ESANN 2012 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 25-27 April 2012, i6doc.com publ., ISBN 978-2-87419-049-0.
Available from http://www.i6doc.com/en/livre/?GCOI=28001100967420.

Datasets	Error			Complexity			
	ET	rET	Lasso	ET	rET	ET/rET	Lasso
Friedman1	0.19587	0.18593	0.282441	29900	885	34	4
Two-norm	0.04177	0.06707	0.033500	4878	540	9	20
SEFTi	0.86159	0.84131	0.988031	39436	2055	19	14

Table 1: Overall assessment (parameters of the Extra-Tree method: $M = 100$; $K = p$; $n_{\min} = 1$ on Friedman1 and Two-norm, $n_{\min} = 10$ on SEFTi).

Effect of the regularization parameter t. The complexity of the regularized ET model is shrunk with the $L1$-norm constraint of equation (2) in a way depending on the value of t. As shown on Figure 1(a), an increase of t decreases the error of rET until $t = 3$, leading to a complexity (Figure 1(b)) of about 900 test nodes. Notice that in general the rET model eventually overfits when t becomes large, although this is not visible on the range of values displayed on Figure 1(a).

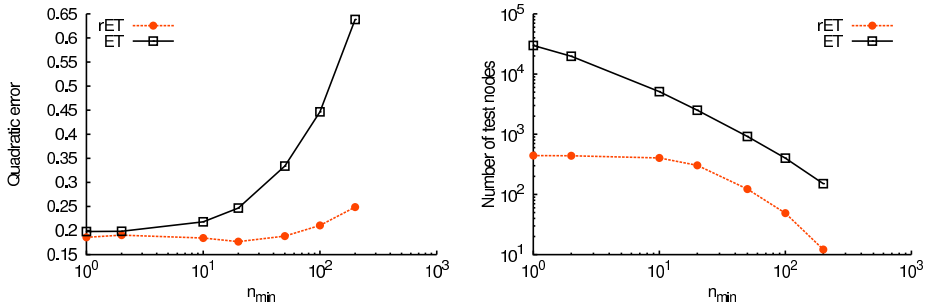


(a) Estimated risk (b) Complexity

Fig. 1: An increase of t decreases the error of rET until t = 3 with drastic pruning ($M = 100$, $K = p = 10$ and $n_{\min} = 1$).

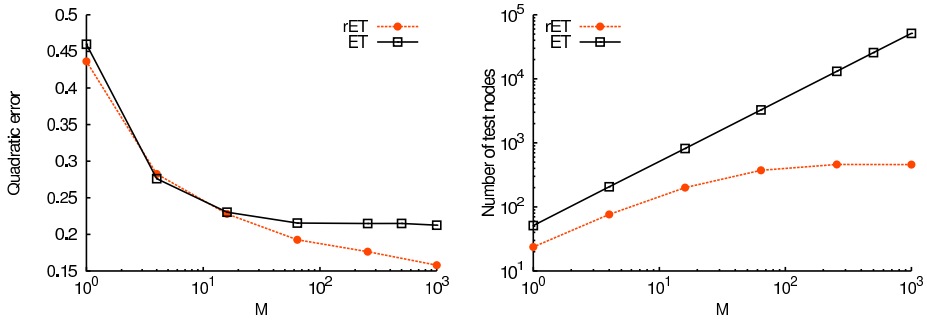
Influence of the Extra-Tree meta parameters $n_{\min}$ and M. The complexity of an ET model grows (linearly) with the size of the ensemble M and is inversely proportional to its pre-pruning parameter $n_{\min}$. Figures 2 show the effect of $n_{\min}$ and Figures 3 the effect of M on both ET and rET models. Interestingly, the accuracy and complexity of the rET model does not depend on the precise value of $n_{\min}$, as long as it is small enough ($n_{\min} \leq 10$, on Figures 2). On the other hand, increasing the value of M beyond the value $M(p)$ where variance reduction has stabilized ($M(p) \simeq 100$ on Figures 3) allows to further improve the accuracy of the rET model without increasing its complexity.

ESANN 2012 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 25-27 April 2012, i6doc.com publ., ISBN 978-2-87419-049-0.
Available from http://www.i6doc.com/en/livre/?GCOI=28001100967420.



(a) Estimated risk (b) Complexity

Fig. 2: The accuracy and complexity of an rET model does not depend on $n_{\min}$, for $n_{\min}$ small enough ($M = 100$, $K = p = 10$ and $t = t^*_{cv}$).



(a) Estimated risk (b) Complexity

Fig. 3: After variance reduction has stabilized ($M \simeq 100$), further increasing M keeps enhancing the accuracy of the rET model without increasing complexity ($n_{\min} = 10$, $K = p = 10$ and $t = t^*_{cv}$).

4 Conclusion

Compression of randomized tree ensembles with $L1$-norm regularization leads to a drastic pruning while preserving accuracy. The complexity of the pruned model does not seem to be directly related to the complexity of the original forest, *i.e.* the number and complexity of each randomized tree, as long as this forest has explored a large enough space of variable interactions.

The strong compressibility of large randomized tree ensemble models suggests that it could be possible to design novel algorithms based on tree-based randomization which would scale in a better way to very high-dimensional input spaces than the existing methods. To achieve this, one open question is how to get the compressed tree ensemble directly, *i.e.* without generating a huge randomized tree ensemble and then pruning it.

Tree-based ensemble models may be interpreted as a lifting of the original input space towards a (randomly generated) high-dimensional discrete and sparse representation, where each induced feature corresponds to the indicator function

ESANN 2012 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 25-27 April 2012, i6doc.com publ., ISBN 978-2-87419-049-0.
Available from http://www.i6doc.com/en/livre/?GCOI=28001100967420.

of a particular tree node, and takes the value 1 for a given observation if this observation reaches this node, and 0 otherwise. The dimension of this representation is on the order of $nM(p)$, but the number s of non-zero components for a given observation is only on the order of $M(p) \log n$. Compressed sensing theory [14] tells us that high-dimensional sparsely representable observations may be compressed by projecting them on a random subspace of dimension proportional to $s \log p$, where p is the original dimension of the observations and $s \ll p$ is the number of non-zero terms in their sparse representation basis. This suggests that one could reduce the space complexity of tree-based method by applying compressed sensing to their original input feature space if its dimension is high, and/or to their induced feature space if $nM(p)$ is too large.

We believe that further developments will benefit from the many results of the compressed sensing field, once the connection between this theory and tree-based ensemble methods is more profoundly established.

Acknowledgements

F. Schnitzler is supported by a F.R.I.A. scholarship. This work was funded by the Biomagnet IUAP network of the Belgian Science Policy Office and the Pascal2 network of excellence of the EC. The scientific responsibility is the authors'.

References

[1] L. Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001. 1

[2] P. Geurts, D. Ernst, and L. Wehenkel. Extremely randomized trees. *Machine Learning*, 63(1):3–42, 2006. 1, 2

[3] P. Geurts. Some enhancements of decision tree bagging. *Principles of Data Mining and Knowledge Discovery*, pages 141–148, 2000. 1

[4] N Meinshausen. Node harvest. *Ann. Appl. Stat.*, 4(4):2049–2072, 2010. 1

[5] J.H. Friedman and B.E. Popescu. Predictive learning via rule ensembles. *The Annals of Applied Statistics*, 2(3):916–954, 2008. 1

[6] N. Meinshausen. Forest garrote. *Electron. J. Statist.*, 3:1288–1304, 2009. 1

[7] Simon Bernard, Laurent Heutte, and Sébastien Adam. On the selection of decision trees in Random Forests. In *Proceedings of the International Joint Conference on Neural Networks*, pages 302–307, France, 2009. 1

[8] G Martínez-Muñoz, D Hernández-Lobato, and A Suárez. An analysis of ensemble pruning techniques based on ordered aggregation. *IEEE Trans. Pattern Anal. Mach. Intell.*, 31:245–259, February 2009. 1

[9] R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 267–288, 1996. 1, 2

[10] T. Hastie, J. Taylor, R. Tibshirani, and G. Walther. Forward stagewise regression and the monotone lasso. *Electronic Journal of Statistics*, 1:1–29, 2007. 2

[11] J.H. Friedman. Multivariate adaptive regression splines. *The Annals of Statistics*, pages 1–67, 1991. 3

[12] L. Breiman. Bias, variance, and arcing classifiers. *Statistics*, 1996. 3

[13] Intel AA&YA. Manufacturing data: Semiconductor tool fault isolation, 11 2008. 3

[14] E.J. Candès and M.B. Wakin. An introduction to compressive sampling. *Signal Processing Magazine, IEEE*, 25(2):21–30, 2008. 6