Stacking Bagged and Boosted Forests for Classification of Noisy and High-Dimensional Data

Raphael R. Campos, Advisor: Marcos André Gonçalves[1]

[1]Departamento de Ciência da Computação
Universidade Federal de Minas Gerais (UFMG)

{rcampos,mgoncalv}@dcc.ufmg.br

Abstract. *Random Forests (RF) are one of the most successful strategies for automated classification tasks. Motivated by the RF success, recently proposed RF-based classification approaches leverage the central RF idea of aggregating a large number of low-correlated trees, which are inherently parallelizable and provide exceptional generalization capabilities. In this context, this work brings several new contributions to this line of research. First, we propose a new RF-based strategy (BERT) that applies the boosting technique in bags of extremely randomized trees. Second, we empirically demonstrate that this new strategy, as well as the recently proposed BROOF and LazyNN_RF classifiers do complement each other, motivating us to stack them to produce an even more effective classifier. Up to our knowledge, this is the first strategy to effectively combine the three main ensemble strategies: stacking, bagging (the cornerstone of RFs) and boosting. Finally, we exploit the efficient and unbiased stacking strategy based on out-of-bag (OOB) samples to considerably speedup the very costly training process of the stacking procedure. Our experiments in several datasets covering two high-dimensional and noisy domains of topic and sentiment classification provide strong evidence in favor of the benefits of our RF-based solutions. We show that BERT is among the top performers in the vast majority of analyzed cases, while retaining the unique benefits of RF classifiers (explainability, parallelization and easiness of parameterization).*

1. Introduction

Organizing and extracting useful information from the enormous quantity of data available nowadays is a vital task for industry and society. By using machine learning (ML) techniques to automatically associate documents with classes, Automatic Text Classification (ATC) provides means to organize information which allows better comprehension and interpretation of the data. Many important applications, such as topic categorization, sentiment analysis, spam filtering, recommender systems, among others, can be effectively and efficiently solved by automatic textual classifiers. Despite the wide applicability of ATC, it brings its own challenges, such as high dimensionality and presence of noise. Properly handling these issues is of great importance to guarantee high effectiveness.

Several ML techniques aimed at tackling the challenging ATC problem have been proposed. In particular, ensembles of classifiers have been shown to excel in this situation [Salles et al. 2015, Dong and Han 2004]. Random Forests (RF) are one of the most successful classifier ensembles in a wide variety of classification tasks

[Fernández-Delgado et al. 2014]. Despite being a classifier with great generalization power, it has been shown that RF models may suffer from overfitting issues [Segal 2004], having its effectiveness degraded in the presence of many irrelevant or noisy attributes — a characteristic of textual classification tasks. More precisely, it has been shown that RF classifiers whose decision trees are grown to their maximum depth are deemed to perform poorly in presence of noisy attributes. Optimistically speaking, these attributes are considerably correlated to one another or are weakly related to the outcome [Salles et al. 2017]. Particularly, when the number of attributes is large, but the fraction of relevant ones is small, random forest models tend to perform poorly. This has to do with the unnecessary variance incurred by the model, as discussed in [Hastie et al. 2009]: the bagged decision trees become plagued by irrelevant or noisy attributes, which introduces unnecessary model complexity (e.g., irrelevant classification paths). Recently, novel RF-based models were proposed to mitigate the overfitting issue by exploiting very distinct strategies, namely, a lazy RF version called LazyNN RF [Salles et al. 2017], and a boosted RF strategy named BROOF [Salles et al. 2015]. Both methods learn classification models focusing on specific sub-regions of the input space, hoping to filter out irrelevant attributes and data— the primary factors that contribute to RF's tendency to overfit.

In our dissertation, we advance the state-of-the-art in ATC by proposing a novel derivation of the RF classifier. More specifically, we propose a new boosted version of RF based on some ideas explored by BROOF: the so-called Boosted Extremely Randomized Trees (BERT) classifier. While BROOF is able to mitigate the overfitting issue faced by RF when applied to high-dimensional noisy data, by avoiding the generation of overly complex trees, it offers limited capability of bias reduction (through the so-called selective out-of-bag based weight update strategy) [Salles et al. 2017]. Thus, bias may still pose as an important factor to contribute to the error rate. To tackle this issue, we introduced another source of randomization in the boosted strategy proposed by [Salles et al. 2015] in order to achieve a better bias-variance tradeoff, by building tree in a more extreme fashion as proposed by [Geurts et al. 2006]. This novel strategy has the following motivations: (i) we expect to avoid overly complex models (and thus mitigate overfitting) through the application of the BROOF-like strategies; (ii) to provide more control over the learner's bias through the additional randomization offered by building extremely randomized trees [Geurts et al. 2006]; and, finally, (iii) to exploit the fact that the extremely randomized trees have shown to be more robust to noise than the RF classifier. Moreover, motivated by the fact that distinct learning methods may complement each other, uncovering specific structures that underlie the input/output relationship of the data at hand, in this work we also propose to exploit the complementary characteristics of the recently proposed RF-based approaches and ours, by stacking them in order to learn an even more effective meta-classifier. As we shall see later, their level of disagreement is high, which motivates our idea. Up to our knowledge, this is the first attempt to combine the three main ensemble strategies: bagging, boosting and stacking. Finally, when stacking classifiers, one usually relies on k fold cross-validation procedures to estimate the *a posteriori* class probabilities for each example, to serve as input for the meta-classifier. Based on these predicted *a posteriori* class distribution estimates, the meta-classifier induces a relationship between these predictions and the true class. However, such estimation strategy may be very costly and sometimes ineffective, since it depends on learning k different models to estimate the probability distributions that serve as input for the stacking procedure.

In order to cope with this problem, we here propose to exploit the efficient and unbiased out-of-bag (OOB) error estimate, an out-of-the-box estimate naturally produced by the bootstrap procedure used in each RF-based learner. Thus, we avoid additional computation efforts to learn a stacked classifier. To summarize, our contributions are fivefold: (1) The proposal of a novel RF-based classifier – BERT – is able to outperform state-of-the-art classifiers; (2) Theproposal of a new stacking classifier that exploits the complementary characteristics of BROOF, LazyNN RF and BERT that is able to outperform all analyzed classification algorithms, including a stacking of traditional state-of-the-art methods, often by large margins; (3) The proposal of new metric for estimating the complementarity among pair of classifiers, named Normalized Degree of Disagrementment. Our proposed metric takes into account the prediction capability of classifiers in order to estimate their complementarity, which can provide a fairer comparison among classifiers with complete distinct accuracy rates. (4) The proposal of a new estimation strategy based on the use of OOB for generating the input for the stacked meta-classifier that substantially reduces the computational effort/runtime of the stacking strategy while retaining its predictive power; (5) Extensive experimentation with 15 datasets in two domains – topic categorization and sentiment analysis; – against several baselines including traditional classifiers (to compare with BERT), several stacking combinations (to compare with the stacking of Forests) and several state-of-the-art stackers (to compare with our OOB-based approach).

1.1. Publications

The contributions of this master's dissertation were published in national and international conferences and proceedings, including:

- Raphael Campos, Sérgio Canuto, Thiago Salles, Clebson C. A. de Sá and Marcos A. Gonçalves.Stacking Bagged and Boosted Forests for Effective Automated Classification. In Pro- ceedings of SIGIR '17, (10 pages - Qualis A1) - the most important worldwide conference on Information Retrieval [Campos et al. 2017]
- Campos, R. R. and Gonçalves, M. A. (2016). Bert: Melhorando classicação de texto com árvores extremamente aleatórias, bagging e boosting. 31st SBBD [Campos and Gonçalves 2016] **(Honorable Mention for Best Short Paper)**.
- R R. Campos, M A. Gonçalves and T. Salles (2016). Quando a Amazônia Encontra a Mata Atlântica: Empilhamento de Florestas para Classificação Efetiva de Texto. KDMiLe 2016 [Campos et al. 2016]

We are also concluding the submission to a top-tier journal with all the dissertation's contributions along with a theoretical analysis of bias and variance of our solutions.

2. Contributions

2.1. Boosted Extremely Randomized Trees

We studied the impact of additional source of randomness in the generalization power of BROOF [Salles et al. 2015]. BROOF combines boosting and bagging by exploiting RFs as "weak learners" in a boosting framework, along with the following strategies: (i) the use the out-of-bag (OOB) error as a less biased error estimation to drive the boosting algorithm; and (ii) to only update the weights of OOB instances during the boosting iterations. In order to answer whether we can improve BROOF generalization power using an additional source of randomness, we proposed the Boosted Extremely Randomized Trees (a.k.a.

BERT). We take the advantage of the aforementioned BROOF-like strategies, however, we build the composing trees in a more extreme fashion. Instead of finding the best split point for each node of the trees, we rely on a more aggressive source of randomization to learn decorrelated trees by combining random cut-point choice and random attribute selection drawn from a randomly chosen subset of features while building the trees, thus guaranteeing reduced tree correlation the ensemble. [Geurts et al. 2006] showed that ensemble of trees built this way can be more robust to noise and may avoid the boosting to get stuck into too noisy hard-to-classify regions of the input space, reducing its bias.

We argue that exploring these strategies through this novel classification framework brings two benefits: it enables us to minimize variance (mitigating the overfitting problem faced by the trees composing the ensemble) and also provide us means to minimize bias, through the additional randomization source, leveraging the framework ability to avoid being stuck on a few hard-to-classify examples.

Our experiments with several datasets (covering text categorization and sentiment analysis domains), comparing with up to nine state-of-the-art classifiers, showed that BERT[1] was among the top performers classifiers in all tested datasets, outperforming the original BROOF in several cases. In all such cases, we showed that the extra source of randomness is an important factor of improvement and bias reduction.

2.2. Normalized Degree of Disagreement

Based upon our results, we noticed that those newly-developed RF-based models excel in both explored text classification tasks. This left us with an unanswered question: can we combine these methods in order to learn a even more effective classifier?. An important step towards the answer was to assess whether these distinct strategies do produce complementary (diverse) information that could be explored to leverage classification effectiveness. To this end, we quantify such complementarity degree by means of the Degree of Disagreement [Kuncheva and Whitaker 2003] observed for a pair of classifiers. Despite being a measurement with values varying from 0 to 1 (1 being the most diverse pair), usually in practice, the values are far from reaching 1. As we showed in the dissertation, this is due to the lower and upper bound of the degree of disagreement determined by the generalization capabilities of the pair of classifiers. This issue can make us unable to determine whether or not a pair of models complement one another. Thus, we proposed the Normalized Degree of Disagreement that scales the degree of disagreement by the maximal and minimal degree given the accuracy of the pair of models, defined as: $Dis_{i,j}^{norm} = \frac{Dis_{i,j} - Dis_{i,j}^{\min}}{Dis_{i,j}^{\max} - Dis_{i,j}^{\min}}$. We demonstrated that the minimal and maximal degree of disagreement between any pair of classifiers with accuracy rates R_i and R_j are defined as $Dis_{i,j}^{\min} = R_i + R_j - 2\min(R_i, R_j)$ and $Dis_{i,j}^{\max} = \min(R_i + R_j, 2 - R_i - R_j)$, respectively.

With such derivations in place, we showed that RF-based models complement each other. This gaves us some evidence that learning methods, such as BROOF, LazyNN_RF and BERT do have some complementary information that can potentially be explored in order to come up with more effective learners. This was the main motivation for our novel strategy to stack RF based classifiers that, besides producing highly effective meta-learners, also enjoys a significantly reduced runtime.

[1] implementation available on: https://github.com/raphaelcampos/stacking-bagged-boosted-forests

2.3. Stacking RF-based Models

We also proposed and studied an efficient way of stacking bagging-based classifiers. When stacking classifiers, one usually relies on k fold cross-validation procedures to estimate the *a posteriori* class probabilities for each example, to serve as input for the meta-classifier [Wolpert 1992]. Based on these predicted *a posteriori* class distribution estimates, the meta-classifier induces a relationship between these predictions and the true class. However, such estimation strategy may be very costly and sometimes ineffective, since it depends on learning k different models to estimate the probability distributions that serve as input for the stacking procedure. In order to cope with this problem, we rely on the out-of-bag samples produced by the bootstraping performed by bagging-based classifiers, such as Random Forests, in order to estimate the *a posteriori* probability distributions for the training samples, thus producing the meta attributes to be fed to the stacked classifier. Since this information is promptly generated at training time, our proposed meta-learner can be built with negligible additional computational effort.

Recall that the bootstrap procedure generates samples $\mathbb{D}_{\textit{boot}}$ comprising of approximately $1 - e^{-1} \approx 63\%$ of the original training set $\mathbb{D}_{\textit{train}}$, with the remaining 36% samples being the so-called out-of-bag samples [Hastie et al. 2009]. In the bagging training process, this procedure is repeated in order to produce several distinct training sets $\mathbb{D}^j_{\textit{boot}}$ for building the ensemble composing trees h_j. Thus, we here propose to use $\mathbb{D}^j_{\textit{oob}} = \mathbb{D}_{\textit{train}} \setminus \mathbb{D}^j_{\textit{boot}}$ to estimate the ensemble class probability distribution at a point $x \in \mathbb{D}_{\textit{train}}$ to be used as meta attributes to train a stacked classifier. This comes at a very low cost, since the meta attributes can be efficiently computed during the training stage of bagged learners, without the needs to perform costly estimation strategies, such as cross-validation. Therefore, let M be the number of bootstrap iterations, $\mathbb{D}^j_{\textit{boot}}|_{j=1}^{M}$ be the bootstrap samples and $h_j|_{j=1}^{M}$ the classifiers trained with the corresponding bootstrap samples. We compute the ensemble OOB probability distribution estimates $p^{oob}(\mathbb{C}|x)$ for each instance $x \in \mathbb{D}_{\textit{train}}$ as: $p^{oob}(\mathbb{C}|x) = \frac{\sum_{j=1}^{M} p^{h_j}(\mathbb{C}|x) I[x \in \mathbb{D}^j_{oob}]}{\sum_{j=1}^{M} I[x \in \mathbb{D}^j_{oob}]}$, where I denotes an indicator function that returns 1 when the m-th classifier did not use x as training instance, 0 otherwise, and $p^{h_j}(\mathbb{C}|x)$ is the class *a posteriori* distribution estimated by h_j. In other words, the OOB *a posterori* class probability distribution $p^{oob}(\mathbb{C}|x)$ is assessed by averaging the class probability distributions $p^{h_j}(\mathbb{C}|x)$, estimated by each individual tree that was built without using x as training sample.

Our experimental results show that stacking only the recently proposed RF-based classifiers and BERT using our OOB-based strategy is not only significantly faster than recently proposed stacking strategies (up to six times) but also much more effective, with gains up to 21% and 17% on MacroF$_1$ and MicroF$_1$, respectively, over the best base method, and of 5% and 6% over a stacking of traditional methods, performing no worse than a complete stacking of methods at a much lower computational effort.

3. Conclusion and Future Work

In this master dissertation, we propose BERT - a boosted version of the extremely randomized trees classifier -that leverages the learner's capability to minimize bias while maintaining high predictive power by properly reducing variance. As our experimental analysis reveal, our proposal enjoys top-notch classification effectiveness, being among the top

performers in the vast majority of experimented datasets. We also propose to stack the explored RF-based classifiers in order to exploit the complementarities observed among those classifiers. In here, we also rely on the out-of-bag samples obtained through bootstrapping the training set when learning the forests to avoid the cost of cross-validation. We show that such novel stacking approach is not only able to provide state-of-the-art classification effectiveness, but also at a significantly lower runtime. As future work, we plan to investigate the benefit of out-of-bag error estimate applied to a more sophisticated early stop strategy. We also intend to take fully advantage from parallelizable potential of the Random Forests (Extra-Trees) built at each iteration of BROOF (BERT). Furthermore, we will focus on weighted sampling for selecting the subset of candidate features as a way to avoid noise. This approach may considerably speedup the process of building the trees in high-dimensional data leading to more compact trees since the weighted sampling increases the likelihood of selecting more discriminative features at each split.

Referências

Campos, R., Canuto, S., Salles, T., de Sá, C. C. A., and Gonçalves, M. A. (2017). Stacking bagged and boosted forests for effective automated classification. In *Proc. of the 40th Int, ACM SIGIR Conf. on Research and Development in Information Retrieval, SIGIR*.

Campos, R. R. and Gonçalves, M. A. (2016). Bert: Melhorando classicação de texto com Árvores extremamente aleatórias, bagging e boosting. In *Proc. of the 31st Brazilian Symposium on Databases, 2016*.

Campos, R. R., Gonçalves, M. A., and Salles, T. C. (2016). Quando a amazônia encontra a mata atlântica: Empilhamento de florestas para classificação efetiva de texto. In *4th Symp. on Knowledge Discovery, Mining and Learning*.

Dong, Y.-S. and Han, K.-S. (2004). A comparison of several ensemble methods for text categorization. In *Services Computing, 2004. (SCC 2004)*, pages 419–422.

Fernández-Delgado, M., Cernadas, E., Barro, S., and Amorim, D. (2014). Do we need hundreds of classifiers to solve real world classification problems? *J. Mach. Learn. Res.*, 15(1):3133–3181.

Geurts, P., Ernst, D., and Wehenkel, L. (2006). Extremely randomized trees. *Mach. Learn.*, 63(1):3–42.

Hastie, T., Tibshirani, R., and Friedman, J. H. (2009). *The Elements of Statistical Learning*. Springer.

Kuncheva, L. I. and Whitaker, C. J. (2003). Measures of diversity in classifier ensembles and their relationship with the ensemble accuracy. *Mach. Learn.*, 51(2):181–207.

Salles, T., Gonçalves, M., and Rocha, L. (2017). Phd dissertation: Random forest based classifiers for classification tasks with noisy data. Federal University of Minas Gerais.

Salles, T., Gonçalves, M., Rodrigues, V., and Rocha, L. (2015). Broof: Exploiting out-of-bag errors, boosting and random forests for effective automated classification. In *SIGIR'15*.

Segal, M. R. (2004). Machine learning benchmarks and random forest regression. Technical report, University of California.

Wolpert, D. H. (1992). Stacked generalization. *Neural Networks*, 5:241–259.