An Empirical Methodology to Analyze the Behavior of Bagging

Fábio Pinto, Carlos Soares, and João Mendes-Moreira

INESC TEC/Faculdade de Engenharia, Universidade do Porto
Rua Dr. Roberto Frias, s/n
Porto, Portugal 4200-465
fhpinto@inescporto.pt csoares@fe.up.pt jmoreira@fe.up.pt

Abstract. In this paper we propose and apply a methodology to study the relationship between the performance of bagging and the characteristics of the bootstrap samples. The methodology consists of 1) an extensive set of experiments to estimate the empirical distribution of performance of the population of all possible ensembles that can be created with those bootstraps and 2) a metalearning approach to analyze that distribution based on characteristics of the bootstrap samples and their relationship with the complete training set. Given the large size of the population of all ensembles, we empirically show that it is possible to apply the methodology to a sample. We applied the methodology to 53 classification datasets for ensembles of 20 and 100 models. Our results show that diversity is crucial for an important bootstrap and we show evidence of a metric that can measure diversity without any learning process involved. We also found evidence that the best bootstraps have a predictive power very similar to the one presented by the training set using naive models.

Keywords: Ensemble Learning, Bagging, Diversity, Metalearning.

1 Introduction

Bagging is an ensemble learning technique that allows to generate multiple predictive models and aggregate their output to provide a final prediction [1]. Typically, the aggregation function is the mean (if the outcome is a quantitative variable) or the mode (if the outcome is a qualitative variable). The models are built by applying a learning algorithm to bootstrap replicates of the learning set. Empirical studies show that bagging is able to reduce the error in comparison with single models and is very competitive with other ensemble learning techniques [2].

In this paper, we propose and apply a methodology to study the performance of the bagging algorithm. We investigate the reasons that affect the influence of a bootstrap (and corresponding model) in the space of sub-ensembles. For that, we compute specific bootstrap characteristics. These measures are then compared

X. Luo, J.X. Yu, and Z. Li (Eds.): ADMA 2014, LNAI 8933, pp. 199–212, 2014.

with the importance of a bootstrap[1] on the predictive performance of ensembles that include the model generated by applying a learning algorithm to it.

Our study is based on Metalearning (MtL) techniques. MtL is the study of principled methods that exploit metaknowledge to obtain efficient models and solutions by adapting machine learning and data mining processes [3]. We aim to gain knowledge about the performance and intrinsic behavior of the bagging algorithm. So, we use MtL in a descriptive approach instead of the more typical predictive framework. For that, we adapted several metafeatures already proposed in the literature [4, 5] and we also introduce some new ones that are very specific of our problem domain.

We tested our proposed methodology empirically by executing experiments with 53 classification datasets collected from the UCI repository [6]. For each dataset, we generated bagged ensembles of decision trees with 20 and 100 models. We were able to generate and test all possible combinations of the ensembles with 20 models. However, for computational reasons, we were forced to sample the number of combinations tested for ensembles with 100 models. We present results that indicate the validity of this sampling procedure. All the insights collected from the metadata describing ensembles with 100 models are compared with the 20 models case. This allowed a validation of our sampling procedure.

Given the descriptive aim of our work, we used standard exploratory data analysis procedures to extract knowledge from the metadata that we generated. The main contributions of this paper are: 1) a methodology based on an extensive experimental procedure and on MtL for empirically studying the performance of bagging; 2) new metafeatures that characterize the relationship between bootstrap samples and the complete training data; 3) an exploratory MtL approach using visualization and a statistical method applied to UCI datasets, yielding interesting observations concerning the relationship between the characteristics of the bootstrap sample and the performance of the bagging ensemble.

This paper is organized as follows. Section 2 describes the related work in the field of ensemble learning particularly focused on the bagging algorithm. Section 3 presents the empirical methodology for studying ensembles and a study of the representativeness of the results obtained by sampling from all the possible ensembles with 100 models. Section 4 describes the MtL approach used in this work as well as the metafeatures. In Section 5, we present the descriptive study on the characteristics of a bootstrap and its importance on the predictive performance of an ensemble. Finally, Section 6 concludes the paper with some final remarks and future work.

2 Related Work

Several papers propose theoretical frameworks that provide important insights on the effectiveness and reasons behind the success of bagging. Breiman [1] argued that aggregating can transform good predictors into nearly optimal ones,

[1] We define an important bootstrap as a bootstrap which its correspondent model belongs to the best combinations of tested ensembles in terms of performance.

highlighting however the importance of using unstable learners (small variations in the training set must generate very distinct models [7]).

Friedman [8] related bagging with the bias and variance decomposition of the error. Shortly, the error is split into two components: bias, associated with the intrinsic error of the learner generalization ability; and variance, associated with the error assigned to the variation in the model from one bootstrap to another. In the context of bagging, Friedman claimed that the variance component is reduced (because of the bootstrapping procedure) without changing the bias.

Domingos [9] presented two alternative hypotheses for the success of bagging: although rejecting the possibility of approximation to the optimal procedure of Bayesian model averaging with an appropriate implicit prior probability distribution, he proved that bagging works effectively because it shifts the prior to a more appropriate region of model space. However, Domingos recognized one important fact: none of the above frameworks relate the success of bagging with the domain characteristics.

Friedman and Hall [10] confirmed Breiman's claim by showing that bagging is most successful when used with highly nonlinear estimators such as decision trees and neural networks. In this study they also found evidence that sub-sampling is virtually equivalent to traditional bootstrap sampling. Bühlmann and Yu [11] provided theoretical explanations of the same claim.

Grandvalet [12] provided an interesting study in which he found that bagging equalizes the influence of examples in a predictor. Bootstrapping a dataset implies that fewer examples have a small influence, while the highly influential ones are down-weighted. The author claims that bagging is useless when all examples have the same influence on the original estimate, is harmful when high impact examples improve accuracy, and is otherwise beneficial.

For the ensemble learning literature, it is important to gain understanding of ensembles performance. One way to understand the behavior of learning processes is MtL. Some papers use MtL in a more descriptive manner with the intention of extracting interesting and useful knowledge of a specific domain. Kalousis et al. [13] used MtL for a meta-descriptive symmetrical study in which they found similarities among classification algorithms and datasets. In another domain, Wang et al. [14] published a paper that focuses on rule induction for forecasting method selection by understanding the nature of historical forecasting data. They provide useful rules that rely on metafeatures for suggesting a specific method. Our application of MtL in this paper resembles more these two papers.

3 Empirical Methodology to Characterize Bagging Performance

Formally, an ensemble F gathers a set of predictors of a function f denoted as $\hat{f}_i$. Therefore, $F = \{\hat{f}_i, i = 1, ..., k\}$ where the ensemble predictor is defined as $\hat{f}_f$.

We propose a methodology to empirically analyze the behavior of bagging. Given a set of k bootstrap samples (also referred to in this paper as bootstraps,

for simplicity), we estimate the empirical distribution of performance of the bagging ensembles that can be generated from all elements of its power set. In other words, we estimate the empirical distribution of performance of all possible ensembles of size 2, 3, ... k that can be generated from those k bootstraps.

This distribution can be used to study the role of a given bootstrap (and respective predictive model $\hat{f}_i$) in the performance of $2^k - 1$ possible ensembles, as done in this paper. Additionally, the distribution can be used to analyze the joint relationship between the bootstrap samples in each ensemble and its performance.

It is easy to understand that is impossible to execute the complete set of experiments for ensembles with a realistically large size, such as k=100, given that the number of combinations to test is $2^k - 1$. Therefore, the only possibility is to estimate the distribution of the performance of all ensembles that can be generated with the set of k bootstraps by sampling from its power set. To investigate the validity of this approach, we carried out the following study.

3.1 Estimating the Distribution of Performance by Sampling from the Power Set of Bootstraps

To validate our methodology based on sampling, we executed the full methodology with k=20 and then we studied the impact of sampling. Based on these results, we extrapolate our findings for k=100. We used the Kullback-Leibler Divergence [15] (KLD) to measure the difference between the probability distributions P and Q, defined as $D_{KL}(P||Q) = \sum_n P_n log_2 \left(\frac{P_n}{Q_n}\right)$ where P is the results obtained by testing $2^k - 1$ combinations of k models and Q a sample of those results. Since the KLD measure is not symmetric, we averaged the divergences, $then D_{KL} = \frac{D_{KL}(P||Q)+D_{KL}(Q||P)}{2}$. Given that this experiment implies a large component of randomness, we executed each sampling procedure 100 times and we averaged the values obtained.

In the first experiment, for each dataset, we progressively increased the sampling proportion and systematically computed the KLD between the sample and the population with k=20. Figure 1 shows, as expected, that as the sampling proportion increases, the divergence between the samples and respective population decreases. One can see that for most of the datasets the fall of the curve is rather fast. Figure 2 shows the same result but the values for the 53 datasets are averaged for each sampling proportion. Again, as expected, the standard deviation and the mean KLD decreases as the proportion of sampling increases.

To assess the hypothesis that increasing the number of models in an ensemble changes the sampling results, we repeated the experiment for ensembles with different k values, from 10 to 19. Figure 3 shows a slight increase in the divergence between the samples of equal proportion and respective populations as k increases. This result is expected given that the introduction of a new model can possibly change the inter-relations between the models and therefore affect the performance of some subsets of models. However, all the curves[2] present a very

[2] Estimated using a Local Polynomial Regression (LOESS).

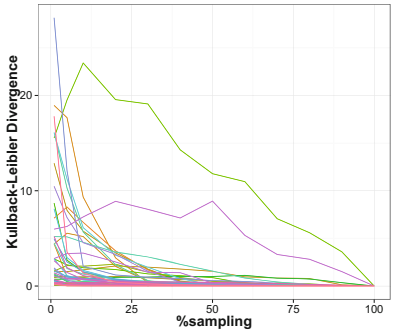

Fig. 1. KLD between % of sample and population. Each line represents a different dataset.

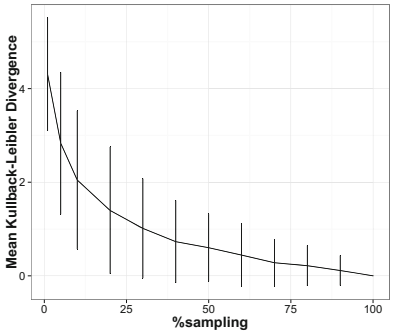

Fig. 2. Mean KLD (and standard deviation, through vertical lines) between % of sample and population

similar pattern. This is indicative that a similar curve could be assumed for an ensemble with k=100.

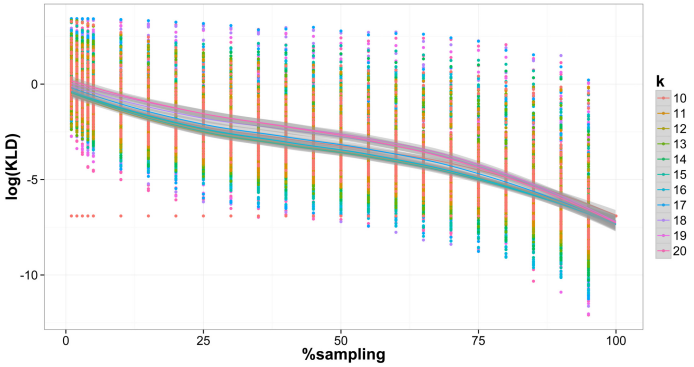

Fig. 3. Sampling and Kullback-Leibler Divergence, averaged for all datasets

3.2 Discussion

Although the evidence showed previously gives us confidence in the sampling variant of our methodology, we still lack sensitivity on the KLD measure to be able to interpret the values of this experiment more reliably. It is difficult by just looking to the graphs if we are actually losing significative information by sampling.

Figures 4 and 5 show two density graphs for a 10% sample and the corresponding complete population. The first concerns the *dis* dataset. One can see that

even for a very large divergence (30.89), the distribution of the sample is very similar to the population distribution. The second graph concerns the *acetylation* dataset, which has a lower divergence (0.23). Most of the datasets show similar values of divergence between their samples and respective populations. This is indicative that we can sample the performance of an ensemble with k=100 and proceed our study.

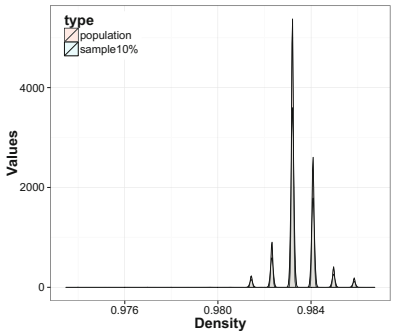

Fig. 4. Density plot for a 10 % sample and population of the *dis* dataset. The KLD between this sample and population is 30.89.

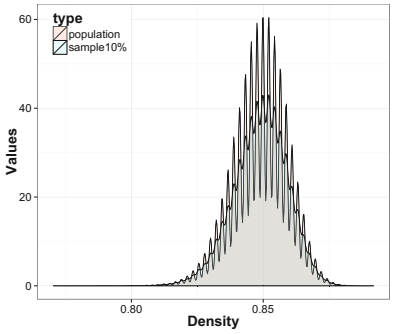

Fig. 5. Density plot for a 10 % sample and population of the *acetylation* dataset. The KLD between this sample and population is 0.23.

The shape of the density graphs is also an interesting result. Both graphs presented a very peculiar pattern of multiple peaks. This is explained by the fact that the bagging performance is a discrete variable. The number of accuracy values that is possible to achieve with all the combinations of a finite set models is limited.

4 Metalearning to Understand Bagging

The methodology presented in Section 3 can be used to provide insights on the types of bootstraps, in terms of how they contribute to the performance of the ensemble. Additionally, it can be combined with a MtL approach to analyze the relationship between the characteristics of the bootstrap sample and the performance of the ensemble.

The main issue in MtL is defining the metafeatures. The most used ones are simple, statistical and information-theoretic metafeatures [3]. In this group we can find the *number of examples* of the dataset, *correlation between numeric attributes* or *class entropy*, to name a few. The use of these kinds of metafeatures provides not only informative data characteristics but also interpretable knowledge about the problems. Other kinds of metafeature are model-based [16]. These capture some characteristic of a model generated by applying a learning

algorithm to a dataset, *i.e.*, the number of leaf nodes of a decision tree. Finally, a metafeature can also be a landmarker [5]. These are generated by making a quick performance estimate of a learning algorithm on a particular dataset.

For this work, we relied on simple, statistical, information-theoretic and landmarker metafeatures. For the first group, we selected several metafeatures already present in the literature which were first used for MtL in the METAL and Statlog projects [3]. We also introduce a new metafeature based on the Jensen-Shannon distance [17] between a bootstrap and the training set. This metafeature aims to measure how different is the bootstrap from the original dataset. It can also be seen as a diversity measure that focuses directly on the bootstrap sample and not on the predictions made by the generated model.

We used two landmarkers: a decision stump and a Naive Bayes classifier. Given the different bias of the algorithms, it is expected that the metafeatures can help capture different patterns. We also used two diversity measures proposed in the ensemble learning literature: the Q-Statistic [18] and Classifier Output Difference [19] (COD) measures. Kuncheva et al. [18] state that the Q-Statistic is the diversity measure with greater potential for providing useful information about ensemble performance. We adapted the Q-Statistic to the specificities of our problem. Kuncheva et al. present it as a metric to measure the diversity of an ensemble. We use it to measure the diversity between the predictions of two models: one generated by applying a learning algorithm to a bootstrap (b) and the other to the original dataset (d). Using such a measure in this study gives a different perspective on its usefulness. Formally, our adapted Q-Statistic is $Q_{b,d} = \frac{N^{bb}N^{dd} - N^{db}N^{bd}}{N^{bb}N^{dd} + N^{db}N^{bd}}$ where each element is formed as in Table 1.

Table 1. Relationship between a pair of classifiers

	f_b**correct**	f_d**correct**
f_b**correct**	N^{bb}	N^{bd}
f_d**correct**	N^{db}	N^{dd}

The COD metric has been proposed as a measure to estimate the potential of combining classifiers

$$C\hat{O}D_T(\hat{f}_b, \hat{f}_d) = \frac{\sum_{x \in Ts} \begin{cases} 1, & \text{if } \hat{f}_b(x) = \hat{f}_d(x) \\ 0, & \text{otherwise} \end{cases}}{|Ts|}$$

in which Ts is test or validation set.

Lee and Giraud-Carrier [20] published a paper on unsupervised MtL in which they study the application of several diversity measures for ensemble learning as a distance function for clustering learning algorithms. In their experiments, only one measure, COD, presents results that indicate that it can be a good measure for this kind of task. This is indicative that the metric can also be useful in our problem.

In summary, the metafeatures used for this work are: *number of examples* of a bootstrap, *number of attributes*, *proportion of symbolic attributes*, *proportion of missing values*, *proportion of numeric attributes with outliers*, *class entropy*, *average entropy between symbolic attributes*, *average mutual information between symbolic attributes and the class*, *average mutual information between pairs of symbolic attributes*, *average absolute correlation between numeric attributes*, *average absolute skewness between numeric attributes*, *average kurtosis between numeric attributes*, *canonical correlation* of the most discriminating single linear combination of numeric attributes and the class distribution, *Jensen-Shannon distance* between the dataset and bootstrap, decision stump *landmarker*, Naive Bayes *landmarker*, Q-Statistic and COD.

The experiments that we carried with the UCI datasets allowed to collect results from the performance of the bagging algorithm in very distinct learning problems. Given that our goal is to understand the importance of each model (and respective bootstrap) in the ensemble space, we need to aggregate the results obtained for each one of them and compute an estimate of importance.

We adapted the measure NDCG [21] (Normalized Discounted Cumulative Gain) to form our metatarget. We consider the performance of the ensembles (in decreasing order) to which the bootstrap k belongs, for each dataset, as $acc_{1,d}$, $acc_{1,d}$, ..., $acc_{n,d}$ where n represents an ensemble and d a dataset. Therefore, for each bootstrap k of the dataset d, we calculate the respective DCG

$$DCG_{k,d} = \sum_{n=1}^{100} acc_{n,d} + \sum_{101}^{n} \frac{acc_{n,d}}{log_{100} n}$$

and we normalize it by an ideal ranking ($IDCG_d$) in which the best ensembles (testing all bootstraps) for each dataset are selected. Then,

$$NDCG_{k,d} = \frac{DCG_{k,d}}{IDCG_d}$$

In order to allow a more concise exploratory analysis of the metadata, we discretized the metatarget. This process is done using the Fisher-Jenks [22] algorithm. The method was chosen since it is well suited to find the optimal partition into different classes of a continuous variable.

5 What Makes a Good Bootstrap?

Most of the metafeatures described characterize the bootstrap in isolation. For instance, the *class entropy* metafeature focuses on the bootstrap and does not relate it with the original dataset. One exception is the diversity measure that characterizes the difference between a set of predictions from a model learned on a bootstrap and another model learned in the original training set. Furthermore, some metafeatures computed for bootstraps of the same dataset show very similar values. For instance, it is not expected that the class entropy varies significantly across bootstrap samples of the same training set. Additionally, the range

of values of a metafeature for different datasets is expected to be quite different. However, we need metafeatures with values in comparable ranges across datasets to be able to extract useful insights with our MtL approach. In summary, we need to transform the metafeatures in order for them to 1) discriminate between bootstrap samples from the same dataset and 2) be comparable across datasets.

So, we applied one of two simple transformations to each meta-variable: **1**) proportional difference of the metafeature computed for the bootstrap in relation to the metafeature computed for the original training set ($\frac{metafeature_d - metafeature_b}{metafeature_d}$) **2**) proportional difference of the metafeature computed for the bootstrap in relation to the maximum value computed for all the bootstraps of the dataset. Then, it is rescaled in order to keep the natural interpretation of the variables by subtracting ($1 - \frac{Max(metafeature_b) - metafeature_b}{Max(metafeature_b)}$). The first transformation was applied to all the metafeatures except the Jensen-Shannon distance, Q-Statistic and COD. To these metafeatures, since we could not compute them in original training set, we applied the second transformation.

The results of the discretization of the metatarget can be verified in Figures 6 and 7. One can see that the discretized values are grouped in a descending order of the value of the metatarget, as it is desirable. Through the analysis of the results we will mention the concept of importance. We consider that bootstraps of class A are more important than bootstraps of class B or C, therefore, we are interested in understanding the characteristics of important bootstraps.

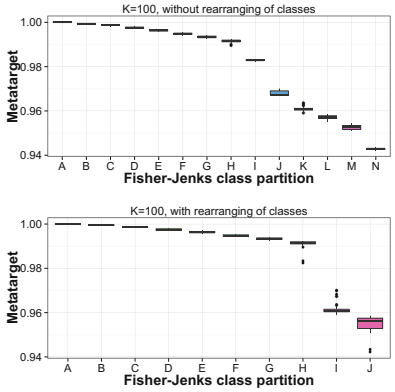

Fig. 6. Boxplot of numeric metatarget (k=100) vs classes found by Fisher-Jenks algorithm

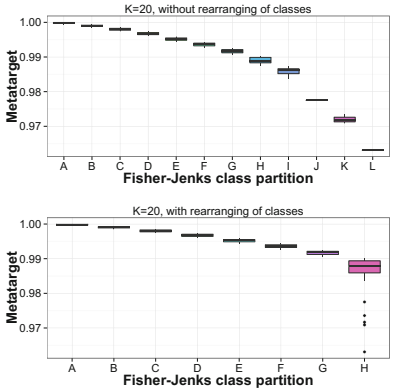

Fig. 7. Boxplot of numeric metatarget (k=20) vs classes found by Fisher-Jenks algorithm

However, some classes group very few observations. It can become problematic to analyze those groups. We decided to merge these classes and reduce the sparsity of the discretization. The graphs at the bottom of Figures 6 and 7 show the boxplots of the metatarget variable after that rearrangement.

5.1 Exploratory Analysis

To assist our analysis, we used Kruskal-Wallis one-way analysis of variance with Wilcoxon pairwise rank sum test as post hoc procedure (0.95 confidence interval) with Holm adjustment method. This analysis was carried to check for significant different medians of the metafeatures among the classes of the metatarget. Figures 8 and 9 show the results of Wilcoxon test for the metafeatures that the Kruskal-Wallis test showed a *p-value* below 0.05. One can see that the metafeatures *avg.symb.pair.mutual.information*, *nb.landmarker* and *q.statistic* are the most discriminative ones. We will focus on metafeatures that are more interesting for the ensemble learning literature and withdraw the analysis of the remaining metafeatures due to space limitations.

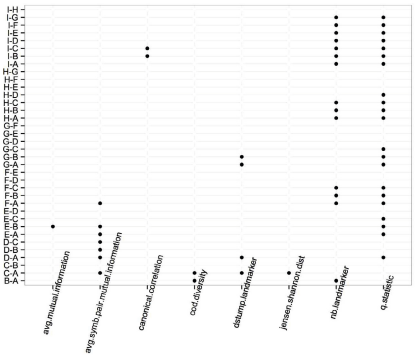

Fig. 8. Pairwise Wilcoxon Rank Sum test for multiple comparison procedures (k=20). Black dot represents a significative difference between the pair of classes.

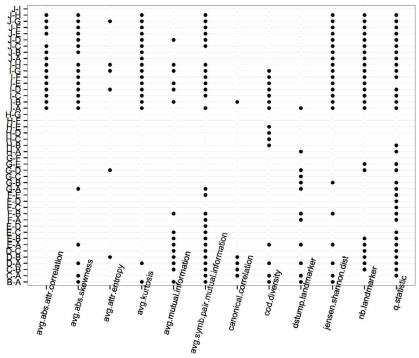

Fig. 9. Pairwise Wilcoxon Rank Sum test for multiple comparison procedures (k=100). Black dot represents a significative difference between the pair of classes.

The Jensen-Shannon distance shows a very interesting pattern that can be verified in Figure 10. One can see that the gradient of the colors associated with each class (in descending order of importance) is reflected in the density distribution graphs. If we compare the distribution of the most important bootstraps (classes A, B, C...) with the less important ones it is clear that, as the Jensen-Shannon distance decreases, the importance of the bootstraps associated with that value also decreases. In other words, bootstraps that are very similar with the original training dataset do not generate a useful model for a bagging ensemble. This is not new for the ensemble learning literature, however, here we measure diversity without any learning process involved. However, this result can not be verified in Figure 11 which represents the metadata with k=20. This can be explained by the fact that since the k=20 experiment generates fewer bootstraps it is harder to find bootstraps with low importance (we can see in

Figure 7 that the range of the metatarget in this experiment is smaller than in the k=100 experiment). However, this remains to be confirmed, which could be done by repeating these experiments for other values of k.

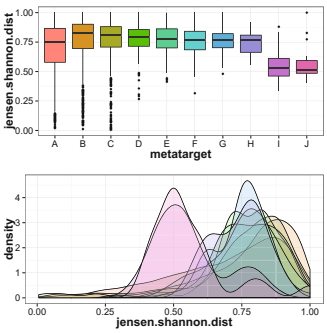

Fig. 10. Boxplot and density distribution of the Jensen-Shannon distance with k=100

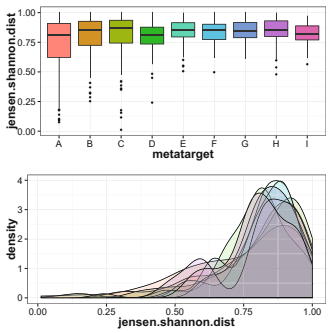

Fig. 11. Boxplot and density distribution of the Jensen-Shannon distance with k=20

Figures 12 and 13 show the density distribution of the diversity measures along the classes of the metatarget. Concerning the Q-Statistic (the bigger, the lesser is the diversity), the results are highly unclear. Although the Wilcoxon test shows that this metafeature has discriminating power, that is not visible graphically. The values of all classes are extremely biased to 1. This may seem contradictory to existing knowledge in the ensemble learning literature, where the Q-Statistic is known to be a good diversity indicator [18].

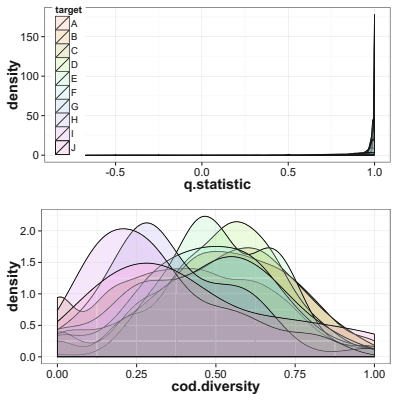

Fig. 12. Density distribution of the metafeatures Q-Statistic and COD for the k=100 experiment

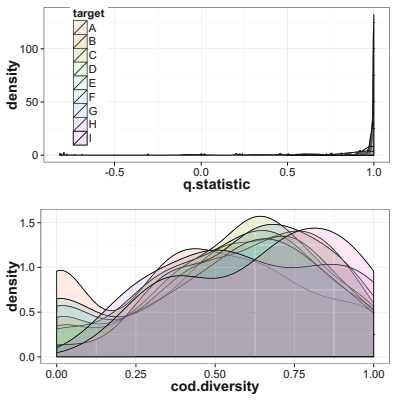

Fig. 13. Density distribution of the metafeatures Q-Statistic and COD for the k=20 experiment

However, we must note that the Q-Statistic is usually computed between models of bootstrap samples while in our case, it is between models of a bootstrap sample and the training set. On the other hand, the COD metric (higher the value, higher is the diversity) shows a very clear direct relationship between diversity and importance of a bootstrap in the k=100 experiment. Again, the result is not confirmed by the k=20 graph. However, we consider this result indicative of the effectiveness of this measure in estimating the potential of combining two classifiers.

Finally, by analyzing the landmarker metafeatures in Figures 14 and 15 we can see interesting patterns. The most prominent one is that important bootstraps have a very similar predictive performance using naive algorithms (such as Naive Bayes and Decision Stump) by comparison against the training set: since we transformed this metafeature as explained previously, a negative value means that the bootstrap has a greater predictive performance than the training set and a positive value the exact opposite. Moreover, we can also see a protuberant peak of the classes that gather the worst bootstraps in the density curves at the left side of the graphs. This indicates that *bad* bootstraps have a superior predictive performance than the training sets.

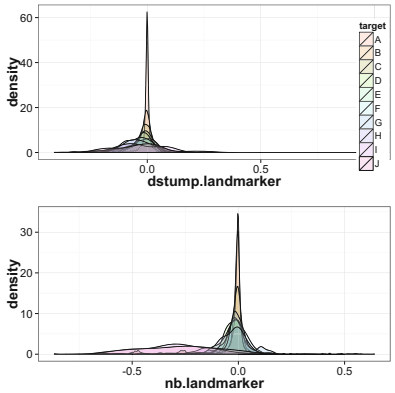

Fig. 14. Density distribution of the landmakers Decision Stump and Naive Bayes for the k=100 experiment

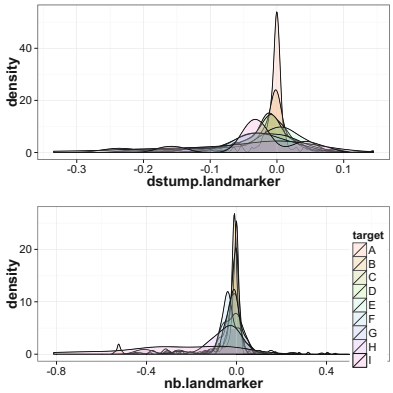

Fig. 15. Density distribution of the landmakers Decision Stump and Naive Bayes for the k=20 experiment

6 Conclusions and Future Work

This paper proposes a methodology based on an extensive experimental procedure and on MtL for empirically studying the performance of an ensemble learning algorithm, more particularly, bagging. We executed experiments with 53 UCI classification datasets using ensembles of decision trees. Initially, we generated 20 models for each dataset and we tested all possible combinations of

those models in the sub-ensemble space. We also executed experiments in which we generated 100 models for each dataset but, due to computational reasons, we were forced to sample the number of combinations tested of the individual models. The results obtained gives us confidence about the effectiveness of the sampling procedure, meaning that it is possible to investigate the distribution of performance of all bagging ensembles obtained with an algorithm by sampling the results. It would be interesting to repeat the experiments with another base learner but we leave that for future work.

To relate the distribution of performance with the characteristics of the bootstrap samples, we adopted an MtL approach. We used several metafeatures proposed in the literature and we introduce three new ones that are very specific of our domain. From our point of view, ensembles are a very promising application of MtL concepts and techniques both to gain a better understanding of their behavior as well as to develop new ensemble methods.

We focused on understanding the characteristics of bootstraps that generate models that are important for the bagging ensemble. We used exploratory data analysis techniques for that goal. Results show interesting patterns that are discriminative of a bootstrap predictive power 1) the bootstrapping procedure should result in a bootstrap sample that is significantly different from the training set, according to the analysis of the Jensen-Shannon distance; 2) the predictions of a model learned from of a bootstrap should be different from the predictions of a model learned from the training, as is known in the ensemble learning literature. However, this observed with the COD metric but not with the Q-Statistic metafeature; 3) the predictive power of a good bootstrap is very similar to the one presented by the training set using naive models.

We plan to extend the work presented in this paper for a predictive MtL approach. The knowledge obtained here can be used to prune a set of bootstraps that can be transformed into an pruned ensemble. It would also be interesting to extend and adapt the methodology proposed in this paper to other ensemble learning algorithms like boosting or random forests. This would bring challenges in the development of the metafeatures in order to deal with probabilistic and random processes.

Acknowledgements. This work is partially funded by FCT/MEC through PIDDAC and ERDF/ON2 within project NORTE-07-0124-FEDER-000059, a project financed by the North Portugal Regional Operational Programme (ON.2 O Novo Norte), under the National Strategic Reference Framework (NSRF), through the European Regional Development Fund (ERDF), and by national funds, through the Portuguese funding agency, Fundação para a Ciência e a Tecnologia (FCT).

References

1. Breiman, L.: Bagging predictors. Machine Learning 24(2), 123–140 (1996)
2. Dietterich, T.G.: An experimental comparison of three methods for constructing ensembles of decision trees: Bagging, boosting, and randomization. Machine Learning 40(2), 139–157 (2000)

3. Brazdil, P., Carrier, C.G., Soares, C., Vilalta, R.: Metalearning: Applications to data mining. Springer (2008)
4. Brazdil, P.B., Soares, C., Da Costa, J.P.: Ranking learning algorithms: Using ibl and meta-learning on accuracy and time results. Machine Learning 50(3), 251–277 (2003)
5. Pfahringer, B., Bensusan, H., Giraud-Carrier, C.: Tell me who can learn you and i can tell you who you are: Landmarking various learning algorithms. In: Proceedings of the 17th ICML, pp. 743–750 (2000)
6. Blake, C., Merz, C.J.: {UCI} repository of machine learning databases (1998)
7. Breiman, L., et al.: Heuristics of instability and stabilization in model selection. The Annals of Statistics 24(6), 2350–2383 (1996)
8. Friedman, J.H.: On bias, variance, 0/1loss, and the curse-of-dimensionality. Data Mining and Knowledge Discovery 1(1), 55–77 (1997)
9. Domingos, P.: Why does bagging work? a bayesian account and its implications. In: KDD, pp. 155–158. Citeseer (1997)
10. Friedman, J.H., Hall, P.: On bagging and nonlinear estimation. Journal of Statistical Planning and Inference 137(3), 669–683 (2007)
11. Büchlmann, P., Yu, B.: Analyzing bagging. Annals of Statistics, 927–961 (2002)
12. Grandvalet, Y.: Bagging equalizes influence. Machine Learning 55(3), 251–270 (2004)
13. Kalousis, A., Gama, J., Hilario, M.: On data and algorithms: Understanding inductive performance. Machine Learning 54(3), 275–312 (2004)
14. Wang, X., Smith-Miles, K., Hyndman, R.: Rule induction for forecasting method selection: Meta-learning the characteristics of univariate time series. Neurocomputing 72(10), 2581–2594 (2009)
15. Kullback, S., Leibler, R.A.: On information and sufficiency. The Annals of Mathematical Statistics, 79–86 (1951)
16. Peng, Y.H., Flach, P.A., Soares, C., Brazdil, P.B.: Improved dataset characterisation for meta-learning. In: Lange, S., Satoh, K., Smith, C.H. (eds.) DS 2002. LNCS, vol. 2534, pp. 141–152. Springer, Heidelberg (2002)
17. Lin, J.: Divergence measures based on the shannon entropy. IEEE Transactions on Information Theory 37(1), 145–151 (1991)
18. Kuncheva, L.I., Whitaker, C.J.: Measures of diversity in classifier ensembles and their relationship with the ensemble accuracy. Machine Learning 51(2), 181–207 (2003)
19. Peterson, A.H., Martinez, T.: Estimating the potential for combining learning models. In: Proceedings of the ICML Workshop on Meta-learning, pp. 68–75 (2005)
20. Lee, J.W., Giraud-Carrier, C.: A metric for unsupervised metalearning. Intelligent Data Analysis 15(6), 827–841 (2011)
21. Järvelin, K., Kekäläinen, J.: Cumulated gain-based evaluation of ir techniques. ACM Transactions on Information Systems 20(4), 422–446 (2002)
22. Fisher, W.D.: On grouping for maximum homogeneity. Journal of the American Statistical Association 53(284), 789–798 (1958)