Proceedings of Machine Learning Research 74:129–140, 2017 Poster - LIDTA 2017

Evaluation of Ensemble Methods in Imbalanced Regression Tasks

Nuno Moniz NMMONIZ@INESCPORTO.PT
Paula Branco PAULA.BRANCO@DCC.FC.UP.PT
Luís Torgo LTORGO@DCC.FC.UP.PT
LIAAD-INESC TEC DCC-FCUP, University of Porto
Porto, Portugal

Editors: Luís Torgo, Bartosz Krawczyk, Paula Branco and Nuno Moniz.

Abstract

Ensemble methods are well known for providing an advantage over single models in a large range of data mining and machine learning tasks. Their benefits are commonly associated to the ability of reducing the bias and/or variance in learning tasks. Ensembles have been studied both for classification and regression tasks with uniform domain preferences. However, only for imbalanced classification these methods were thoroughly studied. In this paper we present an empirical study concerning the predictive ability of ensemble methods bagging and boosting in regression tasks, using 20 data sets with imbalanced distributions, and assuming non-uniform domain preferences. Results show that ensemble methods are capable of providing improvements in predictive ability towards under-represented values, and that this improvement influences the predictive ability of models concerning the average behaviour of the data. Results also show that the smaller data sets are prone to larger improvements in predictive accuracy and that no conclusion could be drawn when considering the percentage of rare cases alone.

Keywords: Imbalanced Domain Learning, Utility-based Evaluation, Ensemble Methods, Regression

1. Introduction

The evolution of data mining and machine learning provided tools to address the issue of understanding the relation between variables, providing valuable insight in diverse domains spanning from finance to meteorology. However, such methods allow for more than modelling interplay between variables, such as the prediction of future values in the same domain, i.e. predictive modelling, which will be addressed in this paper. Consider the example of the meteorology domain. Given a set of temperature observations (training set), data mining enables the learning of models which explain the relationship between future and past values of the temperature. By using such models it is possible to attempt the prediction of future values of some target variable (e.g. the temperature). The type of target variable values influence the type of data mining technique and approach employed. For example, in classification tasks the target value is nominal; in regression and time series forecasting tasks the target value is numeric.

To obtain a model many standard learning algorithms can be used, such as Support Vector Machines, Random Forests or Artificial Neural Networks. Also, solutions involving

the combination of models can be applied, such as ensemble methods. These methods train several models using a given learning algorithm to tackle the same problem, combining their outcome (Zhou, 2012) with averaging or voting approaches.

An important issue in some learning tasks is that the prediction of different values of the target variable may not have the same relevance to users. In many cases a user is focused on predicting what is anomalous or rare instead of what is common or standard. This is a well known problem raised by imbalanced data, which is thoroughly discussed by Branco et al. (2016). Data imbalance is defined as the existence of an over-representation of a given class(es) or numeric value interval(s), over another. Also, in many cases, the under-represented class or numeric value interval is the most relevant for the user and a wrongful prediction may be costly. According to Branco et al. (2016) it is the combination of these two factors (skewed distribution and user preferences towards under-represented items) that forms the basis for the tasks of imbalanced learning.

In scenarios of imbalanced domains, standard learning algorithms bias the models toward the more frequent situations, away from the user preference biases, proving to be an ineffective approach and a major source of performance degradation (Chawla et al., 2004). Concerning ensemble methods, its impact has only been studied within the context of classification tasks, e.g. Wallace et al. (2011) and Galar et al. (2012).

In this paper we propose to study the performance of ensemble methods in imbalanced regression tasks. The main goal of this study concerns the interplay between the predictive accuracy of ensemble methods towards *i)* the average behaviour of the data and *ii)* the rare cases in the data. The ensemble methods studied include bootstrap aggregating (Breiman, 1996) (i.e. bagging) and boosting (Schapire, 1990). An extensive experimental evaluation is presented using 20 data sets and a discussion of the results is introduced.

The remainder of this paper is structured as follows. In Section 2 the concept of ensemble methods is introduced and the bagging and boosting techniques described. In Section 3 the task of imbalanced domain learning is formalized. The experimental study is presented in Section 4, followed by a discussion in Section 5. Conclusions are presented in Section 6.

2. Ensemble Methods

Ensembles are methods designed to use several models with the goal of obtaining an improved predictive performance when compared to the use of a single model. The main idea is to train multiple models combining their predictions through a certain mechanism. These methods are typically more successful than single models (Zhou, 2012) and stand out in a diversity of real-world problems and data mining competitions such as the KDD-Cup[1]. Ensembles have been explored for both classification (Dietterich et al., 2000) and regression tasks (Mendes-Moreira et al., 2012).

The term ensemble usually characterizes methods that combine multiple hypothesis generated by the same learning algorithm while the term "multiple classifier systems" refers to the aggregation of a more diverse set of hypothesis that are not obtained by the same algorithm (Ho, 2002). In this paper we focus on ensembles of models from a given learning algorithm.

1. http://www.kdd.org/kdd-cup

A key aspect of ensembles is the generation of diversity among the models while maintaining the consistency with the training set (Galar et al., 2012). The intuition behind this aspect is clear: in order to obtain gains from the combination of models, they must be different from each other. The bias-variance (Ueda and Nakano, 1996) and the ambiguity (Krogh and Vedelsby, 1994) decomposition of the diversity on ensembles is well known and has been theoretically studied in the context of regression tasks.

In this paper we will use two popular representatives of ensembles methods: boosting and bagging. Regarding boosting algorithms (Schapire, 1990) they generally consist of sequentially training a number T of models and combining them for obtaining the final prediction. The models are trained on an adjusted training set for being able to focus more on the examples that the previous models have failed to accurately predict. One of the most successful boosting algorithm is Adaboost (Freund and Schapire, 1997). Regarding the bagging (bootstrap aggregating) algorithm two main step are involved: bootstrap and aggregating. Bootstrap sampling (Efron and Tibshirani, 1993) is a sampling technique that uniformly and with replacement obtains a new set of m examples from a training set originally containing m examples. In a bagging method, T different models are obtained by using T different bootstrap samples of the training set. Then, the predictions of the T models are aggregated by averaging in regression tasks.

The use of ensembles has been studied in the context of imbalanced domains in classification tasks. Research shows that in classification tasks, due to their design which is focused on the model accuracy, these algorithms are not able to solve the class imbalance problem. However, several approaches have been proposed to deal with this problems through the application of ensemble methods (Galar et al., 2012), requiring the original ensemble algorithms to undergo adaptations. The majority of the proposed solutions to address the problem of imbalanced domains using ensembles resort to pre-processing methods when training models. Also, other type of solutions depend on the embedding of costs into the ensemble learning process.

3. Imbalanced Domain Learning

The problem of imbalanced domains occurs in the context of predictive analytics. Predictive tasks main goal is to obtain a model $Y = h(X)$ that approximates an unknown function $Y = f(X)$. In order to achieve this goal a training set is used. This training set $\mathcal{D} = \{\langle \mathbf{x}_i, y_i \rangle\}_{i=1}^{N}$ includes N examples and is composed of a given number of feature variables $\mathbf{x}$ and a target variable Y. When the target variable is nominal we face a classification task and when it is numeric we face a regression task. In this paper we focus on regression tasks.

The problem of learning from imbalanced domains can be described by the conjugation of the two following assertions: i) the user has non-uniform preferences across the target variable domain; and ii) the most important cases for the user are under-represented in the available data. This means that learning from imbalanced domains is a particular class of predictive tasks where the user assigns more importance to the performance of the model approximation $h(X)$ in certain poorly represented ranges of the target variable. Therefore, it is the conjugation of the user preference bias with the under-representation of some target variable ranges that causes the problem. If, for instance, the data available has

under-represented ranges of the target variable but the user preferences are towards well represented cases, then we do not have a problem of imbalanced domains.

The scientific community has been mostly focused on dealing with imbalanced domains in binary classification tasks. Having only two classes to deal with makes the problem easier to approach. For instance, in this case it is evident which is the most important class: the class with fewer examples (usually known as minority class). Tackling multiclass imbalanced problems is more difficult and less solutions exist. For regression tasks, the continuous nature of the target variable adds an extra level of difficulty when developing solutions for this problem because we potentially have an infinite number of values for the target variable. Therefore, only a few solutions have been put forward for addressing imbalanced domains in regression tasks.

To deal with the problem of assigning a non-uniform importance to a continuous target variable, (Torgo and Ribeiro, 2007) and Ribeiro (2011) proposed the notion of a relevance function, represented by $\phi()$. The key idea of the relevance function is to express the user preferences by defining a mapping between the target variable domain and a scale of relevance, $\phi : \mathcal{Y} \rightarrow [0, 1]$. The extreme values of the relevance, zero and one, represent the assignment of a minimum or maximal relevance respectively. However, it is the user responsibility to define the relevance of the target variable values. This definition is domain dependent, and typically it requires the intervention of domain experts. For scenarios where expert knowledge is not available, Ribeiro (2011) proposed an automatic approach for obtaining this function, based on some assumptions regarding what is more usual in this context. To estimate the relevance function $\phi()$ it is assumed that the most extreme and rare cases of the target variable domain are the most relevant for the user. Having a method to estimate the relevance function, we are now able to define two distinct sets: the set of rare cases $\mathcal{D}_R$ and the set of normal cases $\mathcal{D}_N$. To define such sets, the user is required to set a threshold, t_R on the relevance values. This threshold will be used to build set $\mathcal{D}_R$ and $\mathcal{D}_N$ with the higher/lower relevance values as follows: $\mathcal{D}_R = \{\langle \mathbf{x}, y \rangle \in \mathcal{D} : \phi(y) \geq t_R\}$ and $\mathcal{D}_N = \{\langle \mathbf{x}, y \rangle \in \mathcal{D} : \phi(y) < t_R\}$.

4. Experimental Study

The goal of this paper is to provide a first approach to the study of the well-known ensemble methods boosting and bagging, in the context of imbalanced domain regression tasks. As such, in this section an extensive experimental study is detailed, in order to assess the interplay between the predictive performance of models based on ensembles. The main objective of this experimental study is to observe and discuss the relation between the performance of such models concerning its predictive accuracy towards *i)* the average behaviour of the data, and *ii)* the rare (highly relevant) cases in the data.

Given the scope of the study and the mentioned objectives, this study aims at providing answers to the following research questions:

1. What are the differences between ensemble methods and single models concerning their predictive performance when focusing on their average behaviour and their ability to accurately forecast highly relevant (rare) cases?

2. Does a better evaluation concerning the prediction of average behaviour of data result in a better performance towards highly relevant cases?

3. What is the impact of data sets' characteristics, such as size and percentage of highly relevant cases, concerning the predictive accuracy of the approaches?

4.1. Data

We selected 20 regression data sets from different imbalanced domains. Table 1 shows the main characteristics of these data sets. For each data set, we obtained a relevance function using the automatic method proposed by Ribeiro (2011). This automatic method uses the quartiles and inter-quartile range of the target variable distribution for assigning a higher relevance to both high and low extreme values of the target variable[2]. We considered a threshold of 0.9 on the relevance values in all data sets to obtain the sets of rare and normal cases. This allows us to obtain data sets with different percentages of rare cases, ranging between 2.3% and 15.2%. To ensure the reproducibility of our results all the data and code used is available in http://tinyurl.com/y7n96477.

Table 1: Data sets information by descending order of rare cases percentage. (N: nr of cases; $p.total$: nr predictors; $p.nom$: nr nominal predictors; $p.num$: nr numeric predictors; $nRare$: nr. cases with $\phi(y) > 0.9$; $\%Rare$: $100 \times nRare/N$).

Data Set	N	p.total	p.nom	p.num	nRare	% Rare
a3	198	11	3	8	30	15.2
a6	198	11	3	8	28	14.1
a4	198	11	3	8	27	13.6
a7	198	11	3	8	27	13.6
Abalone	4177	8	1	7	564	13.5
a1	198	11	3	8	22	11.1
boston	506	13	0	13	53	10.5
a5	198	11	3	8	18	9.1
availPwr	1802	16	7	9	141	7.8
a2	198	11	3	8	15	7.6
cpuSm	8192	13	0	13	616	7.5
heat	7400	11	3	8	525	7.1
fuelCons	1764	38	12	26	105	6.0
maxTorq	1802	33	13	20	92	5.1
dElev	9517	6	0	6	478	5.0
bank8FM	4499	9	0	9	198	4.4
dAiler	7129	5	0	5	267	3.7
Accel	1732	15	3	12	61	3.5
ConcrStr	1030	8	0	8	36	3.5
airfoild	1503	5	0	5	35	2.3

4.2. Learning Algorithms

All our experiments were carried out in the R environment. We selected the following learning algorithms: regression trees (RPART), extreme gradient boosting (XGBOOST)

2. Further details available in Ribeiro (2011).

and bagging method (BAGGING). The learning algorithms, respective R packages and the used parameter variants are displayed in Table 2.

Table 2: Regression algorithms, parameter variants, and respective R packages used.

Algorithm	Parameter Variants	R package
RPART	$minsplit = \{20, 50, 100, 200\}, cp = \{0.01, 0.05, 0.1\}$	**rpart** (Therneau et al., 2017)
XGBOOST	$eta = \{0.01, 0.05, 0.1\}, maxdepth = \{5, 10, 15\}$, $cst = \{0.2, \ldots, 0.9\}, nrounds = \{25, 50, 100, 200, 500\}$	**xgboost** (Chen et al., 2017)
BAGGING	$minsplit = \{20, 50, 100, 200\}, cp = \{0.01, 0.05, 0.1\}$, $nbags = \{10, 20, 30, 40, 50\}$	**ipred** (Peters and Hothorn, 2015)

4.3. Evaluation Metrics

It is well known that performance evaluation in imbalanced domains requires the use of special purpose metrics (Ribeiro, 2011; Branco et al., 2016). In fact, standard performance assessment metrics are focused on the average behaviour and are not biased towards the users preferences, providing frequently misleading conclusions (Japkowicz, 2013). Therefore, when handling problems with imbalanced domains it is necessary to take into account the issue of performance evaluation. For classification this issue deserved more attention and several solutions to perform evaluation in this context already exist. Still, for regression, only few evaluation metrics were proposed. A solution for this problem in regression is the framework for obtaining the parallel of precision and recall in imbalanced regression tasks that was proposed by Torgo and Ribeiro (2009) and Ribeiro (2011). We will refer to these metrics as $prec^{\phi}$ and rec^{ϕ} to distinguish them from the classification definitions. This framework is integrates key features of precision and recall measures defined for classification and the notion of numeric error necessary in regression. The key idea behind the definition of $prec^{\phi}$ and rec^{ϕ} metrics is related with the consideration of the Utility of predicting a value $\hat{y}$ for a certain true value y. The Utility of a case depends on the relevance function, ϕ, defined for the problem[3]. Having the $prec^{\phi}$ and rec^{ϕ} notions defined for regression it is easy to derive the F_{β}^{ϕ} metric (cf. Equation 1), where β is a parameter that weight the importance given to precision and recall. Regarding special purpose metrics, in this paper, we use the F_1^{ϕ} metric proposed by Branco (2014) that is based on the mentioned framework.

$$F_{\beta}^{\phi} = \frac{(\beta^2 + 1) \cdot prec^{\phi} \cdot rec^{\phi}}{\beta^2 \cdot prec^{\phi} + rec^{\phi}} \tag{1}$$

For comparison purposes, we will also observe the results through a standard evaluation metric. In this paper we will consider the mse (cf. Equation 2). The consideration of both evaluation perspectives (special purpose and standard evaluation) will allow us to observe the relations between both perspectives concerning models of different learning algorithms.

$$mse = \frac{1}{n} \sum_{i=1}^{n} (\hat{y}_i - y_i)^2 \tag{2}$$

3. Further details regarding this framework can be obtained in Ribeiro (2011).

4.4. Results

Using the materials and methods previously described, an experimental evaluation is carried out. Results are estimated by means of 2 repetitions of a 5-fold cross validation process using the best alternative of each learning algorithm concerning each of the evaluation metrics employed: *mse* and F_1^{ϕ}. The outcome of such estimation is presented in Table 3, with data sets ordered by decreasing percentage of rare cases, and results report the average result over all simulations. The best results concerning each of the evaluation metrics used, for each data set, is denoted in bold.

Table 3: Experimental results by data set and learning algorithm for *mse* and F_1^{ϕ} metrics, estimated by means of 2 repetitions of 5-fold cross validation process.

	RPART				XGBOOST				BAGGING			
	Best *mse*		Best F_1^{ϕ}		Best *mse*		Best F_1^{ϕ}		Best *mse*		Best F_1^{ϕ}	
Dataset	mse	F_1^{ϕ}	mse	F_1^{ϕ}	mse	F_1^{ϕ}	mse	F_1^{ϕ}	mse	F_1^{ϕ}	mse	F_1^{ϕ}
a3	46.410	0.000	56.932	0.470	44.074	0.361	51.245	**0.520**	**43.006**	0.422	**43.006**	0.422
a6	129.162	0.324	149.064	0.526	124.964	0.415	127.871	0.550	**122.463**	0.500	123.130	**0.552**
a4	18.081	**0.618**	18.081	**0.618**	18.593	0.347	24.311	0.577	**17.528**	0.506	17.714	0.588
a7	25.381	0.267	26.843	0.341	24.247	0.385	27.355	**0.422**	**24.179**	0.365	25.473	0.415
Abalone	5.945	0.606	5.959	0.610	**4.621**	0.699	4.623	**0.700**	5.154	0.629	5.175	0.630
a1	307.799	0.000	324.070	0.615	277.576	0.531	296.784	**0.632**	**267.866**	0.000	277.324	0.552
boston	20.929	0.842	20.929	0.842	**9.847**	0.886	10.358	**0.889**	16.002	0.852	16.180	0.856
a5	51.166	0.000	55.747	0.290	44.173	0.250	47.395	**0.519**	**44.158**	0.129	45.497	0.194
availPwr	309.624	0.876	309.624	0.876	**27.790**	0.981	28.154	**0.982**	185.892	0.909	186.808	0.910
a2	113.706	0.058	133.500	0.519	**94.278**	0.197	115.956	**0.537**	98.987	0.192	100.010	0.198
cpuSm	27.163	0.500	41.131	0.671	**6.838**	0.506	7.917	0.566	23.511	0.525	40.639	**0.674**
heat	250.070	0.851	250.070	0.851	**0.556**	0.993	0.567	**0.994**	176.466	0.878	176.466	0.878
fuelCons	0.717	0.811	0.717	0.811	**0.141**	0.931	0.153	**0.935**	0.522	0.838	0.524	0.838
maxTorq	997.336	0.882	997.336	0.882	**34.071**	0.987	35.146	**0.988**	656.141	0.914	656.141	0.914
dElev	**0.001**	0.000	**0.001**	0.000	**0.001**	0.684	**0.001**	**0.692**	**0.001**	0.000	**0.001**	0.000
bank8FM	0.004	0.887	0.004	0.887	**0.001**	0.949	**0.001**	**0.950**	0.003	0.895	0.003	0.895
dAiler	**0.001**	0.262	**0.001**	0.320	**0.001**	0.649	**0.001**	**0.663**	**0.001**	0.547	**0.001**	0.571
Accel	1.953	0.784	2.141	0.843	**0.458**	0.948	0.505	**0.953**	1.375	0.901	1.406	0.901
ConcrStr	90.031	0.085	90.395	0.170	**17.447**	0.764	17.448	**0.765**	57.974	0.085	59.606	0.173
airfoild	19.900	0.105	19.900	0.105	**1.774**	0.337	1.859	**0.351**	13.658	0.020	13.731	0.036

An overall analysis of the results shows that the models that are capable of obtaining the best performance concerning the evaluation metrics used are most of the times based on the XGBOOST algorithm. Results also show that the BAGGING algorithm provides the best results concerning the *mse* evaluation metric in data sets with a higher percentage of rare cases, while XGBOOST algorithm achieves the best *mse* results for data sets with higher rarity. Results concerning the RPART algorithm show empirical evidence to support the claim that ensemble methods provide a better approach to predictive modelling in comparison to the use of single models (RPART). In this case, this is verified concerning both the average behaviour of the data, and the prediction of highly relevant cases (with a single exception for data set *a4*).

By comparing the results obtained by the variants of learning algorithms, evidence also shows that the variants that provide the best outcome concerning the *mse* evaluation metric do not provide a corresponding outcome w.r.t. F_1^{ϕ} metric, and vice-versa. This lack of correspondence is observed concerning all learning algorithms used in our experiments.

In order to further understand the difference between the models used in the experimental evaluation, Figures 1 and 2 show the CD (critical difference) diagrams (Demšar,

2006) according to the non-parametric Friedman test. A lower rank represents better performance, and the horizontal lines connecting the methods show the significance of the difference among ranks. Pairs of models not connected with a horizontal line indicate significant ($p-$value < 0.05) difference in their ranks for a given experiment. In the diagrams, the best variant produced by each algorithm concerning the *mse* metric is denoted with the suffix ".v1", and the best variant concerning the F_1^ϕ metric with ".v2".

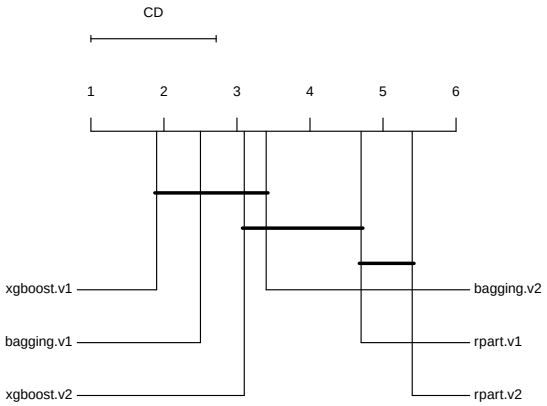

Figure 1: CD diagram for *mse* metric.

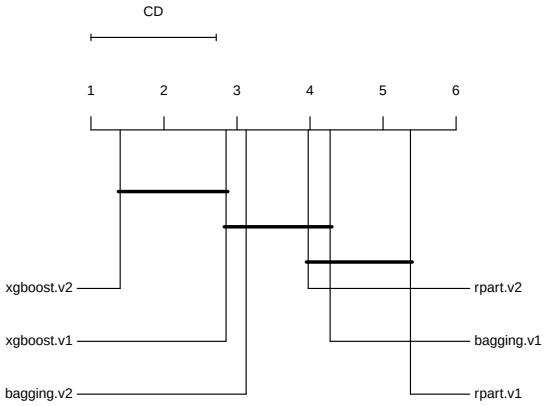

Figure 2: CD diagram for F_1^ϕ metric.

Concerning both the *mse* and F_1^ϕ evaluation metrics, results of the critical difference diagrams confirm the previously stated observations: the algorithm XGBOOST provides the best overall performance, and that its variants (*xgboost.v1* and *xgboost.v2*) present the best results overall for each of the evaluation metrics used for their optimization. Results also point to the lack of statistical significance concerning the variants of each learning algorithm used in the experimental evaluation. In contrast, results show that, in some cases, there is a statistical significance concerning performance between variants of different learning algorithms: *xgboost.v1* provides a significant advantage over both variants using the RPART algorithm w.r.t. the *mse* metric, and *xgboost.v2* over the variants of both the RPART and BAGGING algorithms regarding the F_1^ϕ metric. This last observation is relevant because it shows that if you are in a imbalanced regression context the best way to proceed is to use XGBOOST.

5. Discussion

In this section we further investigate the three issues which are the basis of this paper: *i)* the differences between ensemble methods and single models as to their predictive performance, *ii)* the relation between evaluation results concerning the average behaviour of the data its highly relevant cases, and *iii)* the influence of data characteristics.

Concerning the first issue, in the previous section it is observed that experimental results show how ensemble methods (XGBOOST and BAGGING) provide an advantage over single models (RPART). To understand the magnitude of such advantage between variants of each learning algorithm we applied Wilcoxon signed rank tests to test the hypothesis that the

Table 4: Number of significant ($p-$value < 0.05) wins/ties/losses according to Wilcoxon signed rank tests, concerning the mse and F_1^ϕ evaluation metrics.

	mse			F_1^ϕ		
	Win (Sig)	Tie	Loss (Sig)	Win (Sig)	Tie	Loss (Sig)
RPART.v1	11 (6)	9	0 (0)	0 (0)	9	11 (5)
XGBOOST.v1	19 (9)	1	0 (0)	0 (0)	1	19 (4)
BAGGING.v1	16 (5)	4	0 (0)	0 (0)	4	16 (3)

performance of such variants provide statistically significant improvements in predictive accuracy. Such tests are applied focusing on both mse and F_1^ϕ metrics, in order to infer statistical significance (with $p-$value < 0.05) of the paired differences of the approaches' outcome. Table 4 presents these results. Since the tests only concern two variants for each learning algorithm, results report the outcome for the best variant for the mse metric.

Results show that the the best variants of the learning algorithms w.r.t. the mse evaluation metric always present an improvement over the best variants concerning the F_1^ϕ metric, and that the inverse conclusion also stands. As such, statistical evidence demonstrates that when employing both the single models and ensemble methods studied in this paper, an increased ability to accurately predict the target variable of rare cases is obtained through a trade-off. This means that models which are better at predicting highly relevant values, also provide a worse performance concerning the average behaviour of the data, and vice-versa.

Such conclusion also answers to our second question: the interplay between the predictive performance concerning the average behaviour of the data, and the highly relevant cases. Still, it is important to understand the reason for such outcome. Therefore, Table 5 presents the results for the $prec^\phi$ and rec^ϕ factors used in the F_1^ϕ metric, in all data sets and learning algorithms used. Precision illustrates the fraction of utility gathered by the outcome of prediction models concerning cases which the true target variable is considered to be highly relevant. Recall also illustrates such fraction of utility, but concerning cases which the predicted target variable is considered to be highly relevant.

Table 5: Results of precision ($prec^\phi$) and recall (rec^ϕ).

	RPART				XGBOOST				BAGGING			
	Best mse		Best F_1^ϕ		Best mse		Best F_1^ϕ		Best mse		Best F_1^ϕ	
Dataset	$prec^\phi$	rec^ϕ	$prec^\phi$	rec^ϕ	$prec^\phi$	rec^ϕ	$prec^\phi$	rec^ϕ	$prec^\phi$	rec^ϕ	$prec^\phi$	rec^ϕ
a3	0.000	0.439	0.482	0.462	0.434	0.458	**0.541**	**0.509**	0.478	0.489	0.478	0.489
a6	0.328	0.504	0.535	0.534	0.489	0.486	0.664	0.496	0.555	0.524	**0.577**	**0.542**
a4	0.617	**0.625**	0.617	**0.625**	0.355	0.528	0.623	0.560	0.530	0.605	**0.647**	0.602
a7	0.223	0.386	0.319	0.405	0.375	0.409	**0.468**	**0.418**	0.365	0.394	0.442	0.405
Abalone	0.614	0.600	0.622	0.600	0.753	0.652	**0.754**	**0.654**	0.649	0.612	0.652	0.612
a1	0.000	0.662	0.624	0.675	0.523	0.679	**0.649**	0.680	0.000	0.667	0.569	**0.687**
boston	0.851	0.834	0.851	0.834	0.888	0.883	**0.891**	**0.888**	0.865	0.839	0.874	0.840
a5	0.000	0.538	0.273	0.535	0.269	0.525	**0.558**	0.543	0.147	0.541	0.196	**0.557**
availPwr	0.880	0.873	0.880	0.873	**0.979**	0.983	**0.979**	**0.984**	0.910	0.908	0.911	0.909
a2	0.063	**0.541**	0.549	0.509	0.224	**0.541**	**0.593**	0.519	0.212	**0.541**	0.237	0.540
cpuSm	0.480	0.523	**0.952**	0.518	0.517	0.495	0.574	**0.559**	0.538	0.512	0.942	0.525
heat	0.851	0.850	0.851	0.850	**0.993**	0.993	**0.993**	**0.994**	0.885	0.870	0.885	0.870
fuelCons	0.831	0.794	0.831	0.794	0.934	**0.929**	**0.944**	0.927	0.853	0.823	0.853	0.824
maxTorq	0.876	0.888	0.876	0.888	**0.987**	0.987	**0.987**	**0.988**	0.915	0.913	0.915	0.913
dElev	0.000	0.591	0.000	0.591	0.730	**0.645**	**0.750**	0.644	0.000	0.591	0.000	0.591
bank8FM	0.899	0.876	0.899	0.876	**0.950**	0.949	**0.950**	**0.950**	0.906	0.885	0.906	0.885
dAiler	0.297	0.580	0.357	0.580	0.695	0.610	**0.727**	**0.611**	0.607	0.584	0.637	0.583
accel	0.788	0.868	0.828	0.860	0.947	**0.949**	**0.960**	0.945	0.910	0.892	0.911	0.892
concrStr	0.096	0.650	0.186	0.652	**0.783**	0.746	**0.783**	**0.747**	0.093	0.658	0.187	0.659
airfoild	0.125	0.131	0.125	0.131	0.374	0.317	**0.398**	**0.318**	0.016	0.088	0.030	0.095

The observation of the utility-based metrics of Precision and Recall allows to understand the performance of prediction models concerning their ability to accurately predict the target variable of highly relevant cases. Such type of analysis is vastly applied in the context of imbalanced domain classification tasks, where it is often observed that solutions with increased performance towards highly relevant cases are often prone to false positives, i.e. predicting cases as "rare" (minority) when they are considered "normal" (majority). In the context of regression tasks this would translate as an increase in the rec^{ϕ} metric and a decrease in the $prec^{\phi}$ metric. However, results show that in the great majority of cases, the variants of learning algorithms focusing on optimizing the F_1^{ϕ} metric are capable of increasing both the $prec^{\phi}$ and the rec^{ϕ} metric. This shows that such variants are not only accurately predicting more cases where the target variable is considered to highly relevant, but they are also committing fewer errors regarding such cases.

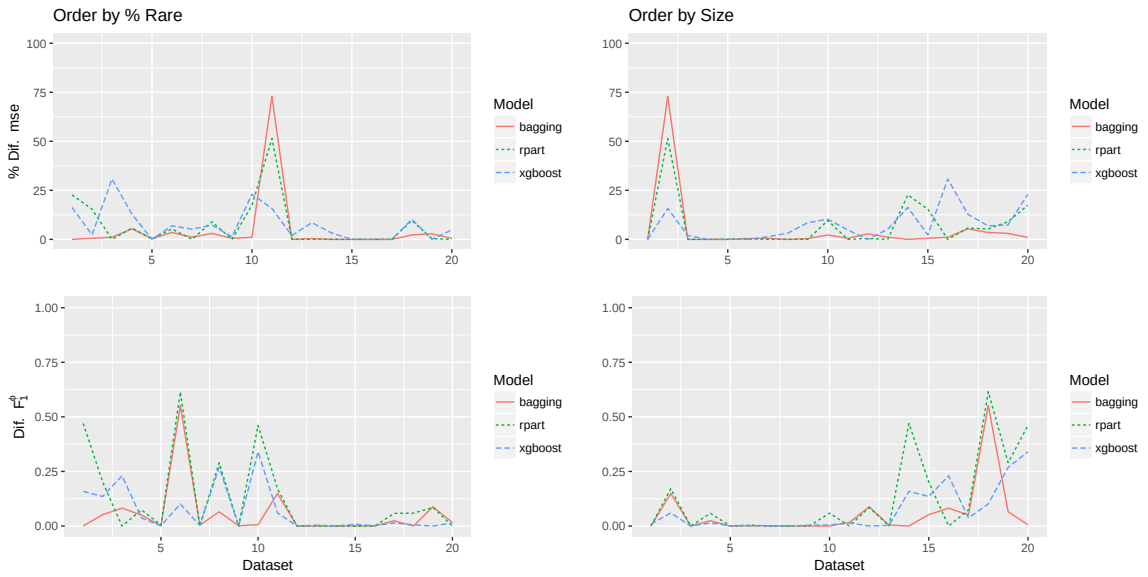

Figure 3: Evolution of the *mse* (percentage) and F_1^{ϕ} (absolute) when comparing the variants optimized by the former with those optimized by the latter for all learning algorithms, with data sets ordered by percentage of rare cases and size.

Concerning the third issue, it is important to understand if the magnitude of improvement in predictive accuracy towards highly relevant cases is related with the two following data characteristics: *i)* the percentage of rare cases, and *ii)* the size of the data set. Figure 3 shows the difference between the variants optimized by *mse* and those optimized by F_1^{ϕ} when ordering the data sets by the two mentioned characteristics. Results concern the percentage of the variation in results concerning the former, and the latter is depicted by their difference. Results are inconclusive when ordering the data sets by decreasing percentage of rare cases. However, for data sets ordered by their size, results show that, for both the *mse* and F_1^{ϕ} metrics, the variants of learning algorithms optimized by the latter (".v2"), show an increasing difference in comparison to the variants optimized by the *mse*

metric. This is interesting, as results show that smaller data sets provide a better setting for improving the ability to accurately predict the target variable of relevant cases.

6. Conclusions

We present an exploratory study concerning the ability of ensemble methods in accurately predicting the target values of highly relevant (rare) cases in imbalanced domain regression tasks. The goal of this study is to address three research questions: *i)* do ensemble methods provide a greater predictive accuracy at predicting rare values in a data set when compared to single models, *ii)* what is the relation between the evaluation of models when focusing on the average behaviour of the data and the highly relevant values of the data, and *iii*) the impact of characteristics in data sets such as the percentage of rare cases and its size.

Results show that, concerning the first question, ensemble methods do provide a greater ability to accurately predict rare values in data sets, with the XGBOOST algorithm providing the best overall approach. As for the second question, results show that it is possible to improve the ability of ensemble models in predicting highly relevant cases, but this comes at a price concerning the evaluation towards the average behaviour of the data. Finally, concerning the third question, results offer evidence that smaller data sets are more prone to improvements concerning the ability to predict highly relevant values in data sets, and that the percentage of rare cases does not explain such predictive ability.

Acknowledgments

This work is financed by the ERDF – European Regional Development Fund through the COMPETE 2020 Programme within project POCI-01-0145-FEDER-006961, and by National Funds through the FCT – Fundação para a Ciência e a Tecnologia (Portuguese Foundation for Science and Technology) as part of project UID/EEA/50014/2013. The work of P. Branco is supported by a PhD scholarship of FCT (PD/BD/105788/2014).

References

Paula Branco. Re-sampling approaches for regression tasks under imbalanced domains. Master's thesis, Dep. Computer Science, Faculty of Sciences - University of Porto, 2014.

Paula Branco, Luís Torgo, and Rita P Ribeiro. A survey of predictive modeling on imbalanced domains. *ACM Computing Surveys (CSUR)*, 49(2):31, 2016.

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

Nitesh V. Chawla, Nathalie Japkowicz, and Aleksander Kotcz. Editorial: Special issue on learning from imbalanced data sets. *SIGKDD Explor. Newsl.*, 6(1):1–6, June 2004.

Tianqi Chen, Tong He, Michael Benesty, Vadim Khotilovich, and Yuan Tang. *xgboost: Extreme Gradient Boosting*, 2017. URL https://CRAN.R-project.org/package=xgboost. R package version 0.6-4.

Janez Demšar. Statistical comparisons of classifiers over multiple data sets. *Journal of Machine learning research*, 7(Jan):1–30, 2006.

Thomas G Dietterich et al. Ensemble methods in machine learning. *Multiple classifier systems*, 1857:1–15, 2000.

Bradley Efron and Robert J. Tibshirani. *An Introduction to the Bootstrap*. Number 57 in Monographs on Statistics and Applied Probability. Chapman & Hall/CRC, 1993.

Yoav Freund and Robert E Schapire. A decision-theoretic generalization of on-line learning and an application to boosting. *J. Comput. Syst. Sci.*, 55(1):119–139, 1997.

Mikel Galar, Alberto Fernandez, Edurne Barrenechea, Humberto Bustince, and Francisco Herrera. A review on ensembles for the class imbalance problem: bagging-, boosting-, and hybrid-based approaches. *IEEE T. Syst. Man. Cy. C*, 42(4):463–484, 2012.

Tin Kam Ho. Multiple classifier combination: Lessons and next steps. In *Hybrid methods in pattern recognition*, pages 171–198. World Scientific, 2002.

Natalie Japkowicz. Assessment metrics for imbalanced learning. In Haibo He and Yunqian Ma, editors, *Imbalanced learning: foundations, algorithms, and applications*. John Wiley & Sons, 2013.

Anders Krogh and Jesper Vedelsby. Neural network ensembles, cross validation and active learning. In *NIPS'94*, pages 231–238. MIT Press, 1994.

Joao Mendes-Moreira, Carlos Soares, Alípio Mário Jorge, and Jorge Freire De Sousa. Ensemble approaches for regression: A survey. *CSUR*, 45(1):10, 2012.

Andrea Peters and Torsten Hothorn. *ipred: Improved Predictors*, 2015. URL https://CRAN.R-project.org/package=ipred. R package version 0.9-5.

Rita P. Ribeiro. *Utility-based Regression*. PhD thesis, Dep. Computer Science, Faculty of Sciences - University of Porto, 2011.

Robert E Schapire. The strength of weak learnability. *MACH*, 5(2):197–227, 1990.

Terry Therneau, Beth Atkinson, and Brian Ripley. *rpart: Recursive Partitioning and Regression Trees*, 2017. URL https://CRAN.R-project.org/package=rpart. R package version 4.1-11.

Luís Torgo and Rita P Ribeiro. Utility-based regression. In *PKDD'07*, pages 597–604. Springer, 2007.

Luís Torgo and Rita P. Ribeiro. Precision and recall in regression. In *DS'09: 12th Int. Conf. on Discovery Science*, pages 332–346. Springer, 2009.

N. Ueda and R. Nakano. Generalization error of ensemble estimators. In *Proceedings of IEEE International Conference on Neural Networks.*, pages 90–95, 1996.

Byron C. Wallace, Kevin Small, Carla Brodley, and Thomas Trikalinos. Class imbalance, redux. In *Proc. of the 11th ICDM*, pages 754–763. IEEE Computer Society, 2011.

Zhi-Hua Zhou. *Ensemble Methods: Foundations and Algorithms*. Chapman & Hall/CRC, 1st edition, 2012.