JMLR: Workshop and Conference Proceedings 39:360–370, 2014 ACML 2014

Ensembles for Time Series Forecasting

Mariana Oliveira MRFO@INESCTEC.PT
Luís Torgo LTORGO@DCC.FC.UP.PT
LIAAD-INESC TEC / DCC-FCUP

Editor: Dinh Phung and Hang Li

Abstract

This paper describes a new type of ensembles that aims at improving the predictive performance of these approaches in time series forecasting. Ensembles are recognised as one of the most successful approaches to prediction tasks. Previous theoretical studies of ensembles have shown that one of the key reasons for this performance is diversity among ensemble members. Several methods exist to generate diversity. The key idea of the work we are presenting here is to propose a new form of diversity generation that explores some specific properties of time series prediction tasks. Our hypothesis is that the resulting ensemble members will be better at addressing different dynamic regimes of time series data. Our large set of experiments confirms that the methods we have explored for generating diversity are able to improve the performance of the equivalent ensembles with standard diversity generation procedures.

Keywords: ensemble methods, time series forecasting, bagging, maximum embed

1. Introduction

Ensembles are known to be among the most competitive forms of solving predictive tasks. Several studies (e.g. Dietterich (2000); Brown and Kuncheva (2010)) have been carried out to understand and explain the reasons for their competitiveness in a wide range of application domains. Diversity among the individual components of ensembles is known to be a key element to generate a successful model. Bagging (Breiman, 1996) is a well known and simple type of ensembles that consists of a large set of standard tree-based models that are grown with the goal of generating a diverse set of models. Diversity in bagging is created through the use of different random bootstrap samples of the original training set to grow each tree. The main idea behind this paper is to propose a variant of bagging where the forms of generating diversity are biased towards specific characteristics of time series forecasting tasks. Namely, we aim at trying to have different forms of handling the diverse dynamic regimes and non-stationarities that are frequently encountered in real world time series. Our working hypothesis is that by using these biased diversity generation methods we will be able to improve the predictive performance of standard bagging on time series forecasting tasks.

The main contribution of this paper is the presentation and experimental analysis of a proposal to improve the predictive accuracy of ensembles on time series forecasting tasks. We describe the general motivation and guidelines for the adaptation of these successful modelling approaches to time series tasks. We present an implementation of our proposal using bagged regression trees and we empirically test its prediction accuracy on a large

set of real world time series. Our results clearly indicate that this is a promising research direction.

In Section 2 we provide a brief description of the tasks being tackled in this paper. Section 3 describes our approach to these tasks using ensembles. The results of an extensive set of experiments are given and discussed in Section 4, while Section 5 analyses some of the related work. Finally, we present the main conclusions of this paper in Section 6 and outline some future research directions.

2. Problem Description

The standard definition of time series forecasting assumes the existence of a set of time-ordered observations of a variable, $y_1, y_2, \cdots, y_t$, where y_i is the value of Y measured at time i, and defines the predictive task as trying to forecast the future values of this variable for time stamps $s > t$. Many variants of this general task exist, including the use of other measured variables as potential predictors of the future values of the target series Y. Still, the general assumption is that there is an unknown function that "maps" the past observations into the future values of Y, i.e. $Y_{t+h} = f(\langle DescriptorsOfThePast \rangle)$, and the learning goal is to approximate this function using some prediction error criterion and a historical record of observed values.

The predictors used for forecasting the future values of Y are usually the most recent observations of Y, as the basic assumption of time series forecasting is that of the existence of some form of correlation between successive observations of the series. This is the approach used on most approaches to time series forecasting, like for instance the well-known ARIMA models (e.g. Chatfield (2013)). This is also the idea of time delay coordinate embedding (Takens, 1981) that is a standard procedure for applying out of the box regression tools to time series forecasting tasks. This strategy assumes that future values of the series are only dependent on a limited number of previous values. In this context, delay coordinate embedding consists in using the k past values of a time series as descriptors of the state of the system at an instant t. If k is appropriate, it's possible to capture the dynamics of the time series from the embed vectors $r_t = \langle y_t, y_{t-1}, ..., y_{t-k} \rangle$. Under this assumption, we can then use any regression tool to obtain a model of the form $Y_{t+h} = f(r_t)$ that specifies the relationship between a set of predictors (described by the embed vector) and the future values of the series.

3. Bagging for Time Series Forecasting

The approaches based on the use of the recent past values of a time series (the embed) as predictors require setting a critical parameter - how many past values to include, i.e. the size of the embed. Setting this parameter is not trivial and it may involve trying different alternative values for the embed size and use some reliable performance estimation process as a means for deciding the "optimal" value. The main drawback of these approaches is the fact that frequently there may not exist *one* single correct answer. In effect, non-stationary series and the occurrence of different regime shifts along time may lead to the best value being clearly time-dependent. This is one of the main motivations for our work. Another being previous work showing that having diversity in ensembles is a key ingredient to boost

their performance. The key idea of our proposal is that of using different sets of predictors (e.g. different embed sizes) within the members of an ensemble to inject some diversity that is related with specific properties of time series tasks. In a nutshell we aim at generating the ensemble members using alternative characterizations of the recent dynamics of the time series. Moreover, by not committing to a single view concerning how the future values depend on the past, we hope to be able to capitalize on having a diverse set of assumptions regarding this dependency within the several models in an ensemble.

There are many possible forms of describing the recent dynamics of a time series through a set of predictor variables. In this paper we present an initial set of proposals for generating different views of this recent dynamics. Namely, we test our hypothesis by having models using: (i) different embed sizes; and (ii) additional features describing summary statistics of the recent values of the series as predictors. Technically, our approach involves having models using less predictors than the ones available in the training set, as well as having other models that use an extended set of predictors by means of some constructed features. More specifically, given a maximum embed size k_{max}, in this paper we will consider the following alternatives:

- E a baseline ensemble where all models are obtained using the maximum embed size, i.e. the previous k_{max} values of the target variable as predictors. This is standard bagging using the maximum embed as predictors. This means that no manipulation of the features available in the provided data is carried out. Diversity in this alternative resorts only to bootstrap samples of the training data as in standard bagging.

- E+S an extension of standard bagging by adding two extra predictors that try to convey extra information on the dynamics of the series, namely μ_Y and σ^2_Y, where the mean and the variance are calculated using the values within the maximum embed, i.e. $\{y_t, y_{t-1}, ..., y_{t-k_{max}}\}$

- DE an ensemble where we have added a new form of diversity on top of random booststrap samples available in standard bagging (E). Namely, we introduced diversity in the embed size used by the different models in the ensemble. One third of the models use the maximum embed, another third uses an embed of $k_{max}/2$ and the last third uses $k_{max}/4$.

- DE+S an ensemble similar to DE but all models will have have the μ_Y and σ^2_Y extra features, although calculated with the respective embed.

- DE±S A small variant on the DE+S alternative, where for each third, half of the models will use the extra statistics, whilst the other half will only use the respective embed.

Table 1 provides a summary of the main characteristics of the bagging variants we are considering and comparing in this paper. Obviously, many more alternatives are possible in terms of trying to generate diversity among ensemble members through strategies related with time series tasks. Our current work only explores some of these alternatives with the aim of trying to improve the predictive performance of ensembles on time series tasks.

The main goal of our proposal is to test the hypothesis that by allowing different views of the recent past observations of a time series within the members of an ensemble, we

Table 1: Summary of the main characteristics of the ensemble variants.

	Embed size	Extra predictors
E	All models use k_{max}.	None.
E+S	All models use k_{max}.	All models use μ_Y and σ^2_Y calculated with the respective embed.
DE	One third of the models use k_{max}, another third uses $k_{max}/2$, and the last third uses $k_{max}/4$.	None.
DE+S	One third of the models use k_{max}, another third uses $k_{max}/2$, and the last third uses $k_{max}/4$.	All models use μ_Y and σ^2_Y calculated with the respective embed.
DE±S	One third of the models use k_{max}, another third uses $k_{max}/2$, and the last third uses $k_{max}/4$.	Half of the models using a certain embed size use μ_Y and σ^2_Y calculated with the respective embed.

will generate some form of biased diversity among them that will be beneficial in terms of predictive performance. Our main contribution is not the features in themselves, but rather their usage within ensemble members as a way of generating diversity related with properties of time series data. To test our hypothesis we have settled on one particular form of ensemble: bagging of regression trees models. In standard bagging diversity among members is obtained by means of using different bootstrap samples of the original training data. For each of these samples an unpruned tree is obtained using all available predictors of the original training set. Our proposal consists of adding an extra form of diversity by varying the used predictors for the different trees according to the schema outlined above. In terms of aggregation of the predictions of the trees to obtain the final prediction of the ensemble we follow the strategy of standard bagging of averaging the predictions of all models. To implement this simple idea we have used the tree-based models available in package **rpart** (Therneau et al., 2014) of the R software environment (R Core Team, 2013), which allows easy replication of our proposal. R code implementing our proposals is freely available at http://www.dcc.fc.up.pt/~ltorgo/ACML2014/.

4. Experimental Evaluation

The main goal of our experimental evaluation is to check the validity of the hypothesis that the variants of bagging we have described in Section 3 will outperform standard bagging on time series forecasting tasks. In this context, our baseline benchmark is a standard bagging implementation using the approach tagged as E in the list given in Section 3. All other four variants will use the same base data (the values of the past k_{max} observations) as training set, but will use it in a different way, e.g. by using only part of it in some models or by using it to generate extra features.

We also compare the ARIMA model, a more standard time series forecasting approach, to the same baseline approach to shed more light onto the overall competitiveness of our proposal. Since ARIMA models usually require a significant parameter tuning effort to

obtain good results, we used the *auto.arima* function available in the R package **forecast** (Hyndman et al., 2014) which automatically searches for an optimal model.

Table 2: Data sets used

ID	Time series	Data source	Data characteristics
1	Temperature	Bike Sharing (Fanaee-T and Gama, 2013)	Daily values from Jan. 1, 2011 to Dec. 31, 2012 (731 values)
2	Humidity		
3	Windspeed		
4	Count of total bike rentals		
5	Temperature		Hourly values from Jan. 1, 2011 to Dec. 31, 2012 (7379 values)
6	Humidity		
7	Windspeed		
8	Count of total bike rentals		
9	Flow of Vatnsdalsa river	Icelandic river (Tong et al., 1985)	Daily values from Jan. 1, 1972 to Dec. 31, 1974 (1095 values)
10	Minimum temperature	Porto weather[1]	Daily values from Jan. 1, 2010 to Dec. 28, 2013 (1457 values)
11	Maximum temperature		
12	Maximum steady wind		
13	Maximum wind gust		
14	Total precipitation		

All five alternative forms of bagging and the ARIMA model were tested on fourteen real world time series obtained from three different data sources as described in Table 2. Each one of these series of data was treated separately from the others in their respective data source (e.g. information on the weather was not used to predict the total number of bike rentals). Moreover, please note that each time series of the Bike Sharing data source is available in two formats (daily and hourly), which were treated as different time series forecasting tasks. All fourteen time series were pre-processed to overcome some well-known issues with this type of data. Specifically, we have created all data sets used in our experiments with the series of the differences between successive values, and not from the original absolute values, in order to avoid trend effects. We have not, however, assumed any shape of these effects, if they exist. The target variable for all tasks was set to the next value of the series of differences.

We have used the standard Mean Squared Error (MSE) as evaluation metric to compare the different approaches. In order to obtain reliable estimates of this metric we have used a Monte Carlo simulation. Time series tasks have an implicit ordering among cases and thus any form of re-sampling will lead to changes of this ordering which is undesirable in terms of reliability of the estimates. In our Monte Carlo experiments we have randomly selected ten points in time within the available time intervals of each task. For each of these ten random points we have used as training set the previous p observations and the following f cases as test set. All approaches were trained and tested using the same exact data to allow for paired comparisons. The size of the training windows (p) was set to 50% of the available data, whilst the test set (f) contained 25% of the cases. The estimates of the MSE

1. Source: http://freemeteo.com.pt

we present are obtained by averaging over these ten random repetitions. Wilcoxon signed rank tests were carried out to test the statistical significance (with $p-$value < 0.05) of the observed paired differences of the proposed approaches against the bagging baseline (E). All experiments were carried out using the experimental infra-structure provided by the R package **performanceEstimation** (Torgo, 2013) and thus can be easily reproduced.

We have repeated our experimental comparisons using four different setups in terms of: (i) number of models in the ensemble (M); and (ii) value of the maximum embed used by the ensembles (k_{max}). Please note that the ARIMA model and its performance does not depend on M and k_{max}. However, since the results we present on this section are relative to the baseline performance which does depend on these parameters, the relative results of the ARIMA model will differ with them.

Table 3 presents the overall results of the paired comparisons of the ARIMA model and of our four variants of bagging against the standard baseline bagging approach that uses an embed of size k_{max} for all M members of the ensemble. The numbers in column "Wins/Losses" are the number of wins and losses of each approach against the baseline, on the fourteen problems. Between parentheses we show how many of these are statistically significant with 95% confidence.

Table 3: Paired comparisons results in format Nr.Wins (Statistically Significant Wins)/ Nr.Losses (Statistically Significant Losses)

M	k_{max}	Variant	Wins/Losses
1020	20	E+S	13 (11) / 1 (1)
		DE	7 (7) / 7 (3)
		DE+S	13 (10) / 1 (0)
		DE±S	14 (12) / 0 (0)
		ARIMA	7 (3) / 7 (4)
	30	E+S	11 (9) / 3 (2)
		DE	10 (6) / 4 (3)
		DE+S	10 (5) / 4 (2)
		DE±S	10 (9) / 4 (2)
		ARIMA	6 (3) / 8 (4)
1500	20	E+S	13 (10) / 1 (1)
		DE	8 (6) / 6 (3)
		DE+S	13 (10) / 1 (0)
		DE±S	14 (12) / 0 (0)
		ARIMA	7 (3) / 7 (4)
	30	E+S	11 (9) / 3 (2)
		DE	9 (7) / 5 (3)
		DE+S	10 (7) / 4 (2)
		DE±S	10 (9) / 4 (2)
		ARIMA	6 (3) / 8 (4)

The results of Table 3 clearly show a positive overall balance of our proposed method for adding time series-specific diversity to bagging. In particular, the DE±S variant achieves remarkable results when $k_{max} = 20$, as it always outperforms standard bagging, most of the times with statistical significance. This is the variant that introduces more variability within the members of the ensemble, which somehow provides further evidence of the advantage of our proposal. Note that the ARIMA model has a much more balanced ratio of wins and losses, achieving a performance apparently very similar to the baseline. Overall, these results are encouraging and provide clear indications of the added value of this research direction even though many more possibilities exist to increase the level of diversity.

Table 4 provides a slightly different perspective of the results of our comparison. We show the average (and standard deviation) rank position of each of the six competitors on all experimental setups. The results once again confirm the validity of our proposal with comparable results obtained by the variants E+S and DE±S.

Table 4: Average and standard deviation of the rank of each method.

M	k_{max}		E	E+S	DE	DE+S	DE±S	ARIMA
1020	20	mean	4.36	**2.00**	4.21	2.29	2.14	3.50
		sd	*0.84*	1.18	0.89	1.07	0.86	2.59
	30	mean	3.93	**2.29**	3.64	2.57	2.57	3.86
		sd	1.44	1.27	1.34	*1.16*	1.28	2.57
1500	20	mean	4.43	**2.00**	4.14	2.29	2.14	3.50
		sd	*0.65*	1.18	1.03	1.07	0.86	2.59
	30	mean	3.86	**2.36**	3.79	2.64	**2.36**	3.86
		sd	1.41	1.28	1.42	1.28	*1.01*	2.57

Table 5 presents the mean (and standard deviation) of the percent difference of MSE compared to standard bagging, i.e., $(MSE_X - MSE_E) \cdot 100/MSE_E$, over all experimental settings. With the exception of the DE variant we observe that on average all our proposals have an overall MSE that is a few percent lower than that of standard bagging. In contrast, the ARIMA model shows a significantly higher average value of MSE than the baseline, which is paired with a much higher standard deviation as well.

Table 5: Percentual difference of MSE with relation to the baseline

M	k_{max}		E+S	DE	DE+S	DE±S	ARIMA
1020	20	mean	**-4.74**	0.22	-4.54	-4.34	36.26
		sd	3.00	3.02	2.83	*2.59*	101.64
	30	mean	-2.30	-0.55	-2.23	**-3.08**	32.63
		sd	6.44	*5.20*	6.96	5.83	110.77
1500	20	mean	**-4.77**	0.27	-4.62	-4.36	36.24
		sd	2.98	3.11	2.80	*2.61*	101.57
	30	mean	-2.28	-0.15	-1.94	**-3.03**	32.76
		sd	6.39	*5.00*	7.01	5.97	110.80

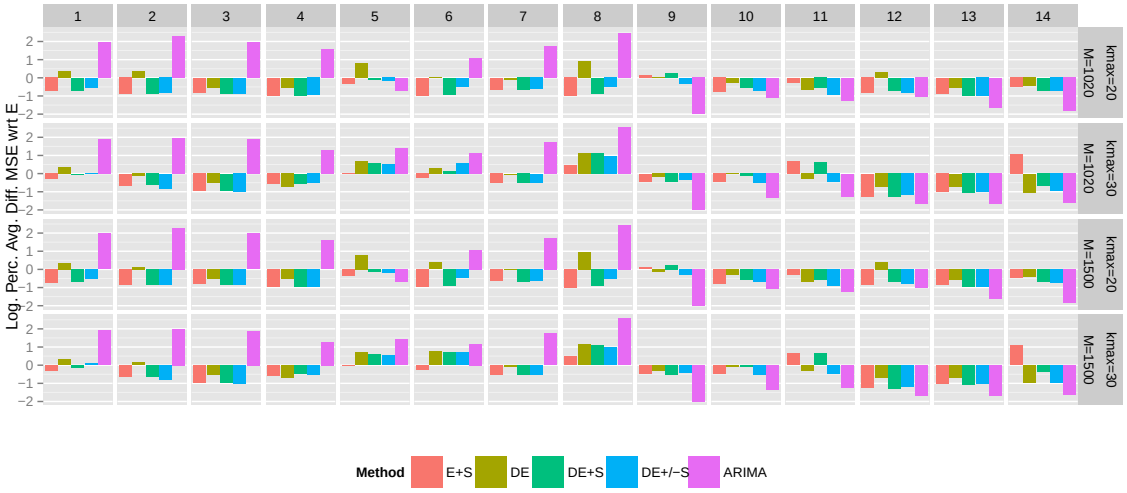

Figure 1: Signed common logarithm of the percent average difference of MSE with relation to the baseline, E, for each time series identified by their respective ID.

Finally, Figure 1 presents more detailed values of the percent difference of MSE with relation to the baseline for each time series (summarized on Table 5). For visualization purposes, a signed common logarithm was applied to the results. The represented metric is, therefore,

$$\operatorname{sgn}(MSE_X - MSE_E) \cdot \log\left(\left|\frac{100 \cdot (MSE_X - MSE_E)}{MSE_E}\right| + 1\right)$$

Once more, our proposals, in general, seem to do well in comparison to standard bagging. The more apparent exceptions to this are the results obtained for $k_{max} = 30$ on time series 5, 6 and 8, with which the ARIMA model seems to struggle as well. The high variance of performance of the ARIMA approach is well illustrated in this figure. Although it achieves a higher decrease in MSE on datasets 9-14, our proposed approaches are also able to perform well on those while almost always beating the results of the ARIMA model on datasets 1-8 with a significant margin.

5. Related Work

Using different predictors on each member of an ensemble is not a novel idea. Random forests (Breiman, 2001) for instance, grow each tree in such a way that for each node a random subset of the features is used to select the best split. Random subspaces (Ho, 1998) is another example of an ensemble method that uses diverse sets of features among the models. This approach consists in randomly selecting subsets (usually of the same size) of the feature space to build each base learner. Contrary to our approach, none of these previous works address time series tasks. Moreover, our subsets of predictors are not

chosen randomly. In effect, our goal is to generate different reasonable forms of describing the recent observations of the target time series, i.e. diversity in the used predictors is guided and not random.

Our approach is also related with recent work on extended space forests for classification (Amasyali and Ersoy, 2013). This work suggests that the addition of new features which are a combination of random pairs of the original features improves the overall accuracy of a decision forest. The improvements obtained for the extended versions of the Bagging algorithm seem to stem from the increase in diversity granted by the extra attributes. In our approach we also propose the extension of the feature space by including summary statistics appropriate for time series.

6. Conclusions

This paper describes an initial attempt at proposing ensembles for time series forecasting tasks. The main motivation of this ongoing work is the observation that handling time series tasks requires several decisions in terms of how we describe the recent dynamics of the observed values of the series. Settling on a single answer to these decisions may be dangerous in real world time series where one frequently observes changes in the dynamic properties of the variable being measured. Ensembles are a well-known answer to this type of problems by taking advantage of diversity among models to reduce both the bias and variance components of the prediction error. Motivated by these observations we have proposed an initial set of forms of injecting diversity into ensembles that takes into account some specific challenges posed by time series data. Namely, we have considered alternative ways of representing the recent observations of the target series among the members of the ensemble. These include the use of different sizes of the embed and also the addition of variables summarising the recent observed values.

We have implemented our ideas in the context of bagging regression trees. The resulting four variants of our proposal have shown a clear advantage over standard bagging in the fourteen real world time series used in this study. The ARIMA model, also compared to standard bagging as a more standard time series forecasting approach, obtained varied performances that depended more on the time series. Although our exploration of this research direction is far from exhaustive, the results we have obtained indicate that this is a promising alternative for time series forecasting tasks.

Future work will include a larger exploration of forms of adding time series specific diversity into ensembles (not only bagging). Namely, we plan to explore: (i) the possibility of changing the amount of past data used by each model (varying training windows); (ii) making the aggregation of the predictions time-dependent; and (iii) use other types of predictor variants.

For the sake of reproducible science that we strongly support, all code and data necessary to replicate all the results shown in this paper are available in the Web page http://www.dcc.fc.up.pt/~ltorgo/ACML2014/. All programs are written in the free and open source R software environment which ensures every one will be able to re-use and reproduce our results.

Acknowledgments

This work is financed by the ERDF European Regional Development Fund through the COMPETE Programme (operational programme for competitiveness) and by National Funds through the FCT (Portuguese Foundation for Science and Technology) within project FCOMP-01-0124-FEDER-037281.

References

M Amasyali and O Ersoy. Classifier ensembles with the extended space forest. 2013.

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

Leo Breiman. Random forests. *Mach. Learn.*, 45(1):5–32, 2001.

Gavin Brown and Ludmila I. Kuncheva. "good" and "bad" diversity in majority vote ensembles. In *Proceeding of Multiple Classifier Systems (MCS 2010)*, pages 1–15, 2010.

C. Chatfield. *The Analysis of Time Series: An Introduction, Sixth Edition*. CRC Press, 2013.

Thomas G. Dietterich. Ensemble methods in machine learning. In *Multiple Classifier Systems*, pages 1–15, 2000.

Hadi Fanaee-T and Joao Gama. Event labeling combining ensemble detectors and background knowledge. *Progress in Artificial Intelligence*, pages 1–15, 2013. ISSN 2192-6352. doi: 10.1007/s13748-013-0040-3. URL http://dx.doi.org/10.1007/s13748-013-0040-3.

Tin Kam Ho. The random subspace method for constructing decision forests. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 20(8):832–844, Aug 1998. ISSN 0162-8828. doi: 10.1109/34.709601.

Rob J Hyndman, with contributions from George Athanasopoulos, Slava Razbash, Drew Schmidt, Zhenyu Zhou, Yousaf Khan, Christoph Bergmeir, and Earo Wang. *forecast: Forecasting functions for time series and linear models*, 2014. URL http://CRAN.R-project.org/package=forecast. R package version 5.6.

R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2013. URL http://www.R-project.org/.

F. Takens. Detecting strange attractors in turbulence. In D. A. Rand and L.-S. Young, editors, *Dynamical Systems and Turbulence*, number 898 in Lecture Notes in Mathematics, page 366381, 1981.

Terry Therneau, Beth Atkinson, and Brian Ripley. *rpart: Recursive Partitioning and Regression Trees*, 2014. URL http://CRAN.R-project.org/package=rpart. R package version 4.1-8.

H. Tong, B. Thanoon, and G. Gudmundsson. Threshold time series modeling of two icelandic riverflow systems1. *JAWRA Journal of the American Water Resources Association*, 21(4):651–662, 1985. ISSN 1752-1688. doi: 10.1111/j.1752-1688.1985.tb05380.x. URL http://dx.doi.org/10.1111/j.1752-1688.1985.tb05380.x.

L. Torgo. *An Infra-Structure for Performance Estimation and Experimental Comparison of Predictive Models*, 2013. URL http://CRAN.R-project.org/web/package=performanceEstimation. R package version 0.1.1.