Bootstrap for model selection: linear approximation of the optimism

G. Simon[1], A. Lendasse[2], M. Verleysen[1, ‡]

Université catholique de Louvain
[1] DICE - Place du Levant 3, B-1348 Louvain-la-Neuve, Belgium,
Phone : +32-10-47-25-40, Fax : +32-10-47-21-80
{gsimon, verleysen}@dice.ucl.ac.be

[2] CESAME - Avenue G. Lemaître 4, B-1348 Louvain-la-Neuve, Belgium,
lendasse@auto.ucl.ac.be

Abstract. The bootstrap resampling method may be efficiently used to estimate the generalization error of nonlinear regression models, as artificial neural networks. Nevertheless, the use of the bootstrap implies a high computational load. In this paper we present a simple procedure to obtain a fast approximation of this generalization error with a reduced computation time. This proposal is based on empirical evidence and included in a suggested simulation procedure.

1 Introduction

A large variety of models may be used to describe processes: linear ones, nonlinear, artificial neural networks, and many others. It is thus necessary to compare the various models (for example with regards to their performances and complexity) and choose the best one. The ranking of the models is made according to some criterion like the generalization error, usually defined as the average error that a model would make on an infinite-size and unknown test set independent from the learning one.

In practice the generalization error can only be estimated, but there exists some methods to provide such an estimation: the AIC or BIC criteria and the like [1], [2], [3] as well as other well-known statistical techniques: the cross-validation and k-fold [3, 6], the leave-one-out [3, 6], the bootstrap [4, 6] and its unbiased extension the .632 bootstrap [4, 6]. The ideas presented in this paper can be applied both to the bootstrap and the .632 bootstrap.

Although these methods are roughly asymptotically equivalent (see for example [5] and [6]), and despite the fact that the use of the bootstrap is not an irrefutable question, it seems that using the bootstrap can be advantageous in many "real world"

‡ G. Simon is funded by the Belgian F.R.I.A. M. Verleysen is Senior Research Associate of the Belgian F.N.R.S. The work of A. Lendasse is supported by the Interuniversity Attraction Poles (IAP), initiated by the Belgian Federal State, Ministry of Sciences, Technologies and Culture. The scientific responsibility rests with the authors.

J. Mira and J.R. Álvarez (Eds.): IWANN 2003, LNCS 2686, pp. 182-189, 2003.

modeling cases (i.e. when the number of samples is limited, the dimension of the space is high, etc.) [6].

But the bootstrap main limitation in practice is the computation time required for assessing an approximation of sufficient reliability (or *accuracy*). A second limitation, in our context of model selection, is the fact that the selected best model is picked up from a set of a priori chosen models, leading to a restricted choice.

In a previous work [7], we have proposed a fast approximation of the generalization error using the bootstrap, based on linear and exponential approximations of the optimism and apparent error (as defined by Efron [4]) respectively. In this paper, we prove experimentally the validity of the linear approximation of the optimism, and show how to use this approximation to perform efficient bootstrap simulations with reasonable computational complexity.

2 Model Selection Using Bootstrap Technique

The fundament of the bootstrap is the plug-in principle [4]. This general principle allows to obtain an estimator of a statistic according to an empirical distribution. In our context of model selection, our statistic of interest is the generalization error. We thus use the bootstrap to estimate the generalization error (or the prediction error in Efron's vocabulary) in order to rank the models and choose the best one.

The bootstrap estimator of the generalization error is computed according to the bootstrap resampling approach. Given an original sample (or data set) x, we generate B new samples, denoted x^b, $1 \leq b \leq B$. The new samples x^b are obtained from the original sample x by drawing with replacement. For each bootstrap sample x^b, we compute a bootstrap estimator of our statistic of interest. The final value is obtained by taking the mean of the estimators over the B bootstrap replications. In the following, we will use the notation $e_{A,B}$ for the error of a model built (learned) on a sample A and tested on a sample B.

In model selection context, Efron defines in [4] the bootstrap estimator of the generalization (prediction) error:

$$\hat{e}_{gen} = e_{app} + optimism\,, \tag{1}$$

where $\hat{e}_{gen}$ is the estimate of the generalization error e_{gen} given by bootstrap, e_{app} is the apparent error (computed on the learning set A), and *optimism* is an estimator of the correction term for the difference between a learning and a generalization error, which in fact aims to approximate the difference of errors obtained on the finite sample x and an (infinite) unknown ideal sample. The optimism is computed according to:

$$optimism = E_B\,[optimism^b]\,, \tag{2}$$

where $E_B[\ ]$ is the statistical expectation computed over the B bootstrap replications and:

$$optimism^b = e_{x^b,x} - e_{x^b,x^b} \ . \tag{3}$$

With our notation, (**1**) becomes:

$$\hat{e}_{gen} = e_{x,x} + E_B\left[e_{x^b,x} - e_{x^b,x^b}\right] \tag{4}$$

In order to approach the theoretical value of the final bootstrap estimation of the generalization error, we can increase the number B of bootstrap replications, but this increases considerably the computation time. We have proposed in [7] a way to reduce this computation load with a limited loss of accuracy.

Note that the .632 bootstrap [4] aims to reduce the slight bias introduced by the *optimism* correction. This bias is due to e_{x^b,x^b} where the error is computed on the same set than the one used for the learning stage. The linear approximation of the *optimism* term, presented in the following, is applicable to the bootstrap and the .632 bootstrap.

3 Framework

Assuming a linear relation between the optimism and the number p of parameters in the model is probably an unexpected hypothesis. Nevertheless, this hypothesis is strengthened by the fact that the general formulation of a structure selection criterion can also be written as

$$\hat{e}_{prediction} = e_{app} + correction \tag{5}$$

where the *correction* term is $2p\sigma/n$ for AIC and $\ln(n)p\sigma/n$ for BIC, with σ the estimated quadratic error on the learning set containing n elements. In AIC, BIC criteria and the like, we can see that the *correction* term is directly proportional to the number of parameter p. Though the apparent error e_{app} is also a function of p, we will focus here on the second term, the *correction*.

Although the *correction* term is computed by bootstrap, and therefore called the *optimism*, its value depends, as the apparent error, on the initialization conditions of the learning process. In practice, a "good" local minimum of a learning error (either on x or on x^b) is obtained by repeating the learning with different initial conditions. Nevertheless, when including this in a bootstrap procedure, the number of learnings is again multiplied by B, resulting in an excessive computation time.

In comparison with the AIC and BIC criteria, we assume that the *correction* term is linearly increasing. Our first goal is then to show experimentally that the *optimism* term is a linear function of p, like a_1p+a_2.

Under this hypothesis, if we compute the value of the *optimism* term for a limited number of models, we can determine (in mean square sense) constants a_1 and a_2. Our second goal is thus, under the linearity hypothesis, to propose a method to reduce the number of tested models and the number of bootstrap replications.

Since the values of a_1 and a_2 result from an experimental procedure, an obvious advantages of our proposal is these parameters are set specifically for each application, avoiding the use of asymptotic results.

4 Methodology

In the experimental results shown below, we used Radial Basis Function Networks (RBFNs) as approximation models. We would like to emphasize on the fact that this choice is made a priori and that the goal is not to compare the results with those that could be obtained with other approximators. The learning procedure to fit the parameters of the model is described in [8], [9].

For the RBFNs models, we consider p in expression a_1p+a_2 as the number of Gaussian units or Gaussian kernels (the total number of parameter in RBFNs is in fact proportional to the number of Gaussian kernels). To observe the linearity, we use the R^2 statistics, also called the square correlation coefficient. The R^2 statistics is here computed between the *optimism* estimated for each model (different values of p) and the linear approximation (a_1p+a_2) of these values. The more this R^2 is close to 1, the most our linear approximation is valid.

Remember that each $optimism^b$, in the context of nonlinear models, is usually the result of several learnings (Q learnings) with different initial conditions. To estimate one value of the *optimism* (i.e. the *optimism* for a specific model complexity p), we should therefore learn QB models, what could be excessive in our context. Now notice that in practice we are not interested in a specific value of the *optimism* but only in the linear approximation a_1p+a_2. A lower accuracy on each value of the *optimism* can thus be balanced by the number of different complexities p, i.e. the number of points (larger than 2) used for the linear approximation.

5 Experimental Results

5.1 Artificial Example

We first illustrate the validity of the linear approximation of the *optimism* described in the previous sections on a toy example. We generate a set of 1000 datas (x, y), with x randomly drawn in [0, 1] and y defined by:

$$y = \sin(5x) + \sin(15x) + \sin(25x) + noise \qquad (6)$$

where *noise* is a uniform random variable in [-0.5, 0.5].

We then use the bootstrap resampling method in a model selection procedure, observing the generalization error corresponding to a specific model characterized by its number p of Gaussian kernels. Figure 1 presents the evolution of our R^2 criterion versus the number B of bootstrap replications. We clearly see that R^2 is getting closer and closer to one while B increases. Fishers' test (with a p-value of $2.2584\ 10^{-11}$) leads to accept the linear hypothesis from $B = 15$.

Since we admit the linear approximation hypothesis, we can go one step further and address the reduction of computation time. We then look to the evolution of a_1 and a_2 in function of B respectively in Figures 2.1 and 2.2. Here again, when B is greater or equal to 15, we have a roughly constant value. Figure 3 shows the graph of the optimism according to the number p of Gaussian kernels in the model. Figure 4 is the graph of the generalization error versus the number of bootstrap replications, where we can see that the "best" model for our toy example has 20 Gaussian kernels. Finally, Figure 5 shows the 1000 learning data and the predictions we got with the selected model.

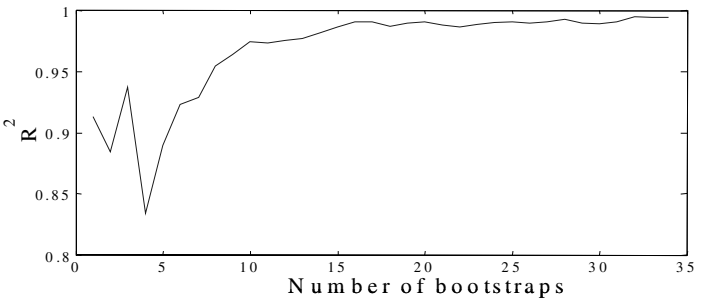

Fig. 1. Toy example: evolution of R^2 versus B

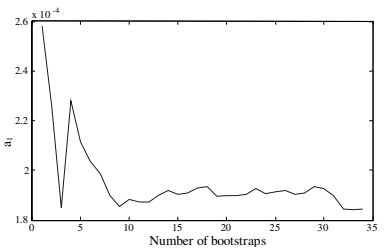

Fig. 2.1 Toy example: evolution of coefficient a_1 versus B

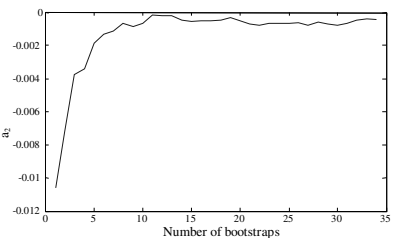

Fig. 2.2. Toy example: evolution of coefficient a_2 versus B

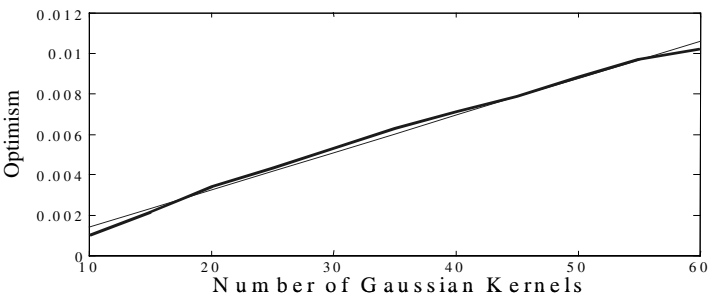

Fig. 3. Toy example: approximation of the *optimism* term (thin line) versus the number of Gaussian kernels with $B = 20$ (thick line : values obtained for the tested models)

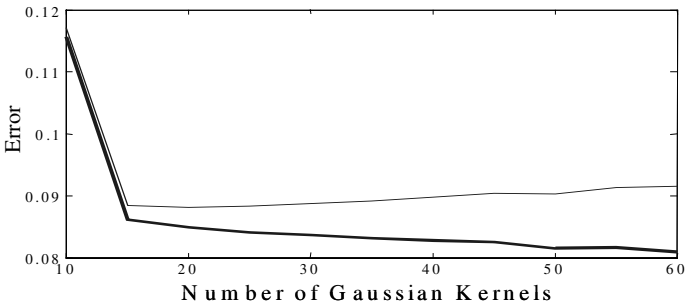

Fig. 4. Toy example: Learning (thick) and generalization (thin) errors versus p

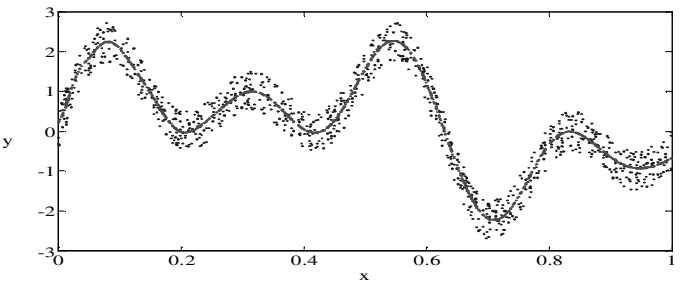

Fig. 5. Toy example: learning data (dots) and predictions (solid line) with the selected model

5.2 Real Data Set (Abalone)

We use the abalone dataset [10] as a second example to validate the linearity hypothesis, with a more realistic (and difficult) approximation problem. Here again, we use 1000 data for the learning. Figure 6 is the evolution of R^2 with respect to B. In this case, Fisher's test p-value is rounded by the computer to 0. Figure 7.1 is the graph of a_1 and figure 7.2 is the evolution of a_2 in function of B. According to these graphs, we suggest to use $B = 40$. Figure 8 shows the reported optimism with respect to p. Figure 9 shows the learning and generalization errors versus B. The minimum corresponding to the "best" model for the Abalone data set has 62 Gaussian kernels.

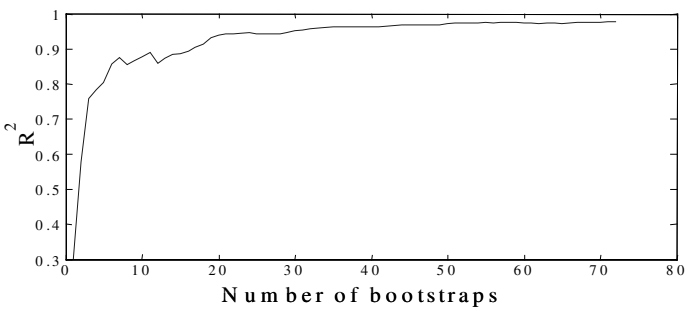

Fig. 6. Abalone: evolution of R^2 versus B

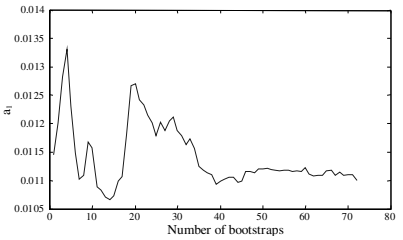

Fig. 7.1 Abalone: evolution of coefficient a_2 versus B

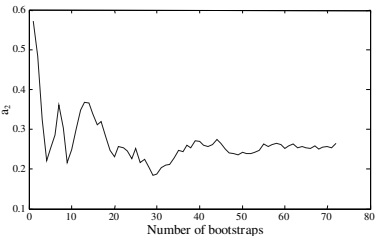

Fig. 7.2 Abalone: evolution of coefficient a_2 versus B

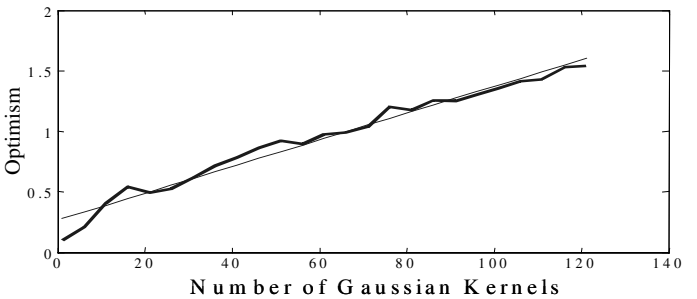

Fig. 8. Abalone: approximation of the *optimism* term (thin) versus the number of Gaussian kernels with $B = 40$ (thick: values obtained for the tested models)

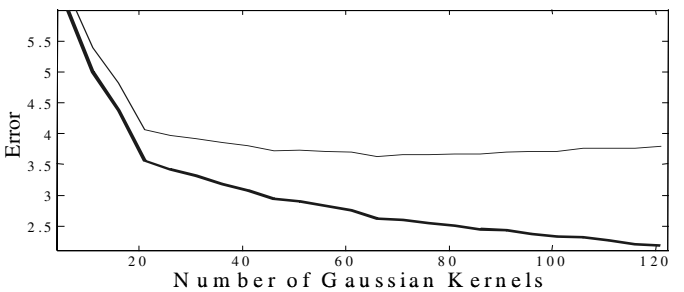

Fig. 9. Abalone: Learning (thick) and generalization (thin) errors versus the number p of Gaussian kernels

6 Conclusion

In this paper we have shown that the *optimism* term of the bootstrap estimator of the prediction error is a linear expression of the number of parameters p.

Furthermore, we illustrate the time saving procedure proposed in [7], enhanced here with the early stop criterion based on the R^2 of the linear approximation. According to the two results shown here and to other ones not illustrated in this paper,

we recommend a conservative value of 50 for the number B of bootstrap replications before stopping the approximation computation.

We would like to emphasize on the fact that the limited loss of accuracy is balanced by a considerable saving in computation load, this last fact being the main disadvantage of the bootstrap resampling procedure in practical situations. This saving is due to the reduced number of tested models and to the limited number of bootstrap replications.

Although this procedure has only been tested in a neural network model selection context, this simple and time saving method could easily be extended to other contexts of nonlinear regression, classification, etc., where computation time and complexity play a role. It can also be applied to other resampling procedures, as the .632 bootstrap.

References

[1] H. Akaike, "*Information theory and an extension of the maximum likelihood principle*", 2nd Int. Symp. on information Theory, 267-81, Budapest, 1973

[2] G. Schwarz, "*Estimating the dimension of a model*", Ann. Stat. 6, 461-464, 1978.

[3] L. Ljung, "*System Identification - Theory for the user*", 2nd ed, Prentice Hall, 1999.

[4] B. Efron, R. J. Tibshirani, "An introduction to the bootstrap", Chapman & Hall, 1993.

[5] M. Stone, "*An asymptotic equivalence of choice of model by cross-validation and Akaike's criterion*", J. Royal. Statist. Soc., B39, 44-7, 1977.

[6] R. Kohavi, "*A study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection*", Proc. of the 14th Int. Joint Conf. on A.I., Vol. 2, Canada, 1995.

[7] G. Simon, A. Lendasse, V. Wertz, M. Verleysen, "Fast approximation of the bootstrap for model selection", accepted for publication in Proc. of ESANN'2003, d-side, Brussels, 2003.

[8] N. Benoudjit, C. Archambeau, A. Lendasse, J. Lee, M. Verleysen, *"Width optimization of the Gaussian kernels in Radial Basis Function Networks"*, Proc. of ESANN'2002, d-side, Brussels, 2002.

[9] M. J. Orr, "*Optimising the Widths of Radial Basis Functions*", in Proc. of Vth Brazilian Symposium on Neural Networks, Belo Horizonte, Brazil, december 1998

[10] W.J. Nash, T.L. Sellers, S.R. Talbot, A.J. Cawthorn and W.B. Ford, "*The Population Biology of Abalone (_Haliotis_ species) in Tasmania. I. Blacklip Abalone (_H. rubra_) from the North Coast and Islands of Bass Strait*", Sea Fisheries Division, Technical Report No. 48, 1994.