

A LATE-STOPPING METHOD FOR OPTIMAL AGGREGATION OF NEURAL NETWORKS

PABLO M. GRANITTO, PABLO F. VERDES, HUGO D. NAVONE and H. ALEJANDRO CECCATTO
*Instituto de Física Rosario, Consejo Nacional de Investigaciones Científicas y Técnicas and
Universidad Nacional de Rosario, Blvd. 27 de Febrero 210 Bis, 2000 Rosario, Argentina
E-mail: {granitto,verdes,navone,ceccatto}@ifir.edu.ar*

Ensembles of artificial neural networks have been used in the last years as classification/regression machines, showing improved generalization capabilities that outperform those of single networks. However, it has been recognized that for aggregation to be effective the individual networks must be as accurate and diverse as possible. An important problem is, then, how to tune the aggregate members in order to have an optimal compromise between these two conflicting conditions. We propose here a simple method for constructing regression/classification ensembles of neural networks that leads to overtrained aggregate members with an adequate balance between accuracy and diversity. The algorithm is favorably tested against other methods recently proposed in the literature, producing an improvement in performance on the standard statistical databases used as benchmarks. In addition, and as a concrete application, we apply our method to the sunspot time series and predict the remainder of the current cycle 23 of solar activity.

1. Introduction

Ensemble techniques have been used recently to improve the generalization capabilities of artificial neural networks (ANNs).¹ The motivation for this procedure is based on the intuitive idea that by combining the outputs of several individual predictors one might improve on the performance of a single generic one. Good ensembles must have accurate but diverse members, which poses the problem of generating a set of ANNs with both reasonably good individual generalization capabilities and independently distributed predictions for the test points.

The diversity of ANNs comes naturally from the inherent data and training process randomness, and also from the intrinsic non-identifiability of the model. On the other hand, there is a trade-off between the ensemble diversity and the generalization capabilities of the individual networks. Some attempts^{2,3} to achieve a good compromise between

these properties include elaborations of bagging⁴ and boosting⁵ techniques.

We provide here a simple way of generating an ANN ensemble with members that have a good compromise between accuracy and diversity. The method essentially amounts to the sequential aggregation of individual predictors where, unlike in standard aggregation techniques which combine individually optimized ANNs,⁶ the learning process of a new member is validated by the overall aggregate prediction performance. That is, the early-stopping method is applied by monitoring the generalization capabilities of the previous-stage aggregate predictor plus the network being currently trained (see Section 3). In this way we retain the simplicity of independent network training and only the validation process becomes slightly more involved, leading in general to some controlled overtraining ("late-stopping") of the individual networks.

We test the proposed algorithm in the regression setting by comparing it against a standard bagging technique adapted from Ref. 4, a simple early-stopping method of individual networks, and the recently-proposed NeuralBAG algorithm (a description of these methods can be found in Ref. 2). For this comparison we use as benchmarks the Ozone, Boston Housing and Friedman#1 statistical databases. In addition, we apply the method to the well-known sunspot time series and compare the results of our algorithm with those of an optimal ensemble averaging of independently-trained ANNs.⁶ The results obtained here for this problem are, to the best of our knowledge, the most accurate ones reported in the vast literature on sunspot prediction.

The organization of this work is the following: In Section 2 we discuss the so called bias/variance dilemma, which provides the theoretical setting for ensemble averaging. In Section 3 we present our method for tuning the individual members of the ensemble and give some insights to understand how this method works. Then, in Section 4 we show empirical evidence of its effectiveness by applying it to the Ozone, Boston Housing and Friedman#1 databases and comparing the results obtained with those of Ref. 2. In Section 5 we consider the sunspot time series and predict the remainder of cycle 23 of solar activity. Finally, in Section 6 we draw some conclusions.

2. The bias/variance dilemma

The theoretical framework for ensemble averaging is based on the bias/variance decomposition of the generalization error.⁷ Let us consider a set of N noisy data pairs $D = \{(t_i, \mathbf{x}_i), i = 1, N\}$, where the vectors \mathbf{x}_i of predictor variables are obtained from some distribution $P(\mathbf{x})$ and the regression targets t_i are generated according to

$$t_i = f(\mathbf{x}_i) + \varepsilon_i. \quad (1)$$

Here f is the true regression and ε is a random noise with zero mean. If we estimate f using an ANN trained on D and obtain a model f_D , the (quadratic) generalization error on a test point (t, \mathbf{x}) averaged over all possible realizations of the data set D (with respect to P and noise ε) can be decomposed as:

$$E[(t - f_D(\mathbf{x}))^2 | D]$$

$$= E[\varepsilon^2 | \varepsilon] + (E[f_D(\mathbf{x}) | D] - f(\mathbf{x}))^2 + E[(f_D(\mathbf{x}) - E[f_D(\mathbf{x}) | D])^2 | D] \quad (2)$$

The first term on the right-hand side is simply the noise variance σ_ε^2 ; the second and third terms are, respectively, the squared bias and variance of the estimation method. From the point of view of a single estimator f_D , we can interpret this equation by saying that a good method should be no biased and have as little variance as possible between different realizations. It is in general believed that the first condition is reasonably well met by ANNs; however, as stated in the introduction, the second one is in general not satisfied since, even for a particular data set D , different training experiments will reach distinct local minima of the error surface (non-identifiability of the ANN model).

A way to take advantage of this apparent weakness of ANNs is to make an aggregate of them. If we rewrite the error decomposition in the form:

$$E[(t - E[f_D(\mathbf{x}) | D])^2 | D] = \text{Bias}^2 + \sigma_\varepsilon^2 = \text{Error} - \text{Variance}, \quad (3)$$

we can reinterpret this equation in the following way: using the average $E[f_D | D]$ as estimator, the generalization error can be reduced if we are able to produce fairly accurate models f_D (small Error) while, at the same time, allowing them to produce the most diverse predictions at every point (large Variance). Of course, there is a trade-off between these two conditions, but finding a good compromise between accuracy and diversity seems particularly feasible for largely unstable methods like ANNs. Several ways to generate an ensemble of models with these characteristics have been discussed in the literature.^{2,6,8} In the next section we propose a new method and give some arguments that suggest why it should be effective; these ideas are later supported by empirical evidence on synthetic and real-world data.

3. Tuning the diversity of ensemble members

As suggested in the previous section, in order to improve the generalization capabilities of the aggregate predictor one must generate accurate but diverse individual networks. This can be accomplished by the following procedure:

Step 1: Generate a training set T_1 by a bootstrap re-sample⁹ from dataset D and a validation set V_1 by

collecting all instances in D that are not included in T_1 . Produce a model f_1 by training a network on T_1 until a minimum $e_{f_1}(V_1)$ of the generalization error on V_1 is reached.

Step 2: Generate new training and validation sets T_2 and V_2 respectively, using the procedure described in Step 1. Produce a model f_2 by training a network until the generalization error on V_2 of the aggregate predictor $\varphi_2 = 1/2(f_1 + f_2)$ reaches a minimum $e_{\varphi_2}(V_2)$. In this step the parameters of model f_1 remain constant and the model f_2 is trained with the usual (quadratic) cost function on T_2 .

Step 3: Iterate the process until an optimal number N_A of models are produced. This optimal number can be estimated by keeping an external validation set or simply from the behavior of $e_{\varphi_n}(V_n)$ as a function of n .

Notice that in the algorithm described above the individual networks are trained in the usual way, but with a late-stopping method based on the current ensemble generalization performance. The method seems to reduce the ensemble generalization error without paying much attention to whether this improvement is related to enhancing ensemble diversity or not. We can see that it actually finds diverse models to reduce the aggregate error as follows. Let us assume that after n iterations we have an aggregate predictor φ_n which produces an average error $e_{\varphi_n}(V_n)$ on the validation set V_n . When we train model f_{n+1} , the average validation error on V_{n+1} of the aggregate predictor φ_{n+1} will be

$$\begin{aligned} e_{\varphi_{n+1}}(V_{n+1}) &= \frac{1}{(n+1)^2} \{e_{f_{n+1}}(V_{n+1}) + n^2 e_{\varphi_n}(V_{n+1}) \\ &+ 2nE[(t - f_{n+1}(\mathbf{x}))(t - \varphi_n(\mathbf{x}))|(t, \mathbf{x}) \in V_{n+1}]\}. \end{aligned} \quad (4)$$

In general we will have $e_{\varphi_{n+1}}(V_{n+1}) < e_{\varphi_n}(V_{n+1})$ (otherwise enlarging φ_n would be useless) and we expect $e_{\varphi_{n+1}}(V_{n+1}) < e_{f_{n+1}}(V_{n+1})$ due to the overtraining of model f_{n+1} (see next section). Then

$$\begin{aligned} E[(t - f_{n+1}(\mathbf{x}))(t - \varphi_n(\mathbf{x}))|(t, \mathbf{x}) \in V_{n+1}] &< e_{\varphi_{n+1}}(V_{n+1}), \end{aligned} \quad (5)$$

which is only possible if f_{n+1} is at least partially anticorrelated to the aggregate φ_n . This analysis shows that at every stage the algorithm is seeking a new diverse model anticorrelated with the current ensemble. In the next section we will show how this heuristic works on real and synthetic data corresponding

to the Ozone, Boston Housing and Friedman#1 statistical databases.

4. Evaluation on benchmark databases

We have used three regression problems to evaluate the algorithm described in the previous section: the real-world Ozone and Boston Housing and the synthetic Friedman#1 data sets. We compared our method against a "Simple" early-stopping method of individual networks, a so called "Benchmark" technique adapted from Ref. 10, and the recently-proposed NeuralBAG algorithm. Descriptions of the three methods can be found in Ref. 2. In order to compare the different methods' performances we used the same training process of individual networks for all of them, changing only the stopping point selection criterion. We also set $N_A = 30$ to allow direct comparison with results in Ref. 2, although this number of networks is not necessarily optimal for our method. All the results quoted below correspond to the average over 50 independent runs of the whole procedure, without discarding any anomalous case. Notice also that the indicated standard deviations only characterize the dispersion in performances due to different realizations of training and test sets; they have no direct relevance in comparing the average performances for different methods since in each run all of them use the same data.

• Ozone

The Ozone data correspond to meteorological information (humidity, temperature, etc.) related to the maximum daily ozone (regression target) at a location in the Los Angeles basin. Removing missing values one is left with 330 training vectors, containing 8 inputs and 1 target output in each one. The data set can be downloaded by ftp (to [ftp.stat.berkeley.edu/pub/users/breiman](ftp://ftp.stat.berkeley.edu/pub/users/breiman)) from the Department of Statistics, University of California at Berkeley.

Like in Ref. 2, we have considered ANN architectures with 5 hidden units trained by the backpropagation rule with learning rate 0.1 and momentum 0.2. Furthermore, we performed the same (random) splitting of the data into training, external validation (for the Benchmark technique) and test sets containing, respectively, 125, 125 and 80 patterns.

From Table 1 we can see that the average mean-squared error obtained with our method is smaller than the corresponding errors produced by the Simple and NeuralBAG algorithms, and only slightly bigger than that of the Benchmark algorithm (which uses information contained on a dataset twice as large as the other methods). Furthermore, our method performs better than the Simple one in 36 out of 50 runs.

• Boston Housing

This data set consists of 506 training vectors, with 11 input variables and 1 target output. The inputs are mainly socioeconomic information from census tracts on the greater Boston area and the output is the median housing price in the tract. It can be downloaded from the UCI Machine Learning Repository (<ftp://ics.uci.edu/pub/machine-learning-databases>).

We followed again the experimental setting in Ref. 2, and considered 200 training examples and 106 data points for the test set, with the remaining 200 examples used as an external validation set for the Benchmark technique. Networks with 5 hidden units were trained by the backpropagation rule with learning rate 0.1 and momentum 0.2.

In this case our method has the best performance, even better than the Benchmark technique. Furthermore, it outperforms the Simple method in 35 out of the 50 runs.

• Friedman#1

The Friedman#1 synthetic data set corresponds to training vectors with 10 input and 1 output variables generated according to

$$t = 10 \sin((x_1 x_2) + 20(x_3 - 0.5)^2 + 10x_4 + 5x_5 + \eta), \quad (6)$$

where η is Gaussian noise with distribution $\mathcal{N}(t, \infty)$ and x_1, \dots, x_{10} are uniformly distributed over the interval $[0, 1]$. Notice that x_6, \dots, x_{10} do not enter in the definition of t and are only included to check the prediction method's ability to ignore these inputs.

We generated 1400 sample vectors and randomly split the data into training, external validation and test sets containing, respectively, 200, 200 and 1000 patterns. We considered ANNs with 6 hidden units, with a learning rate of 0.1 and a momentum of 0.9 for the backpropagation rule.

The corresponding average mean-squared errors on the test set are given in Table 1, where we again include the performances of the other methods considered in Ref. 2 for comparison. In this case our algorithm produced the best performance and also a more than 20% reduction on the standard deviation of ensemble errors. In particular, it outperforms Simple in 44 out of the 50 runs.

Finally, it is of interest to consider the average number of training epochs that individual networks have to be trained before being aggregated to the ensemble. Table 2 gives these figures for the different methods considered and shows that both NeuralBAG and the algorithm here proposed lead to an important overtraining (late stopping) compared to the Simple and Benchmark methods (which essentially correspond to the standard early-stopping method of single networks).

Table 1. Mean-squared test errors averaged over 50 runs corresponding to five different algorithms for ensemble learning. The Simple, Benchmark and NBAG algorithms are described in Ref. 2; the results for Single correspond to the average performance of a single ANN. The standard deviations only characterize the performance fluctuations due to different realizations of training and test sets.

Method	Ozone	Boston	Friedman#1
Simple	18.91 ± 3.21	14.78 ± 6.97	2.49 ± 0.45
Benchmark	18.48 ± 3.03	14.50 ± 6.70	2.43 ± 0.38
NBAG	18.72 ± 3.22	14.96 ± 7.40	2.50 ± 0.48
Single	21.55 ± 4.15	19.95 ± 8.87	4.82 ± 1.54
This Work	18.59 ± 3.20	14.46 ± 6.89	2.32 ± 0.35

Table 2. Average number of training epochs of individual networks required by the different aggregation methods considered.

Dataset	Simple	Benchmark	NBAG	This Work
Ozone	2318	2173	3194	3927
Boston	3879	3722	4955	5460
Friedman#1	7640	6935	14692	17510

5. Application: The sunspot time series

Sunspots are dark blotches on the sun whose mechanism for appearance is not exactly known. Yearly averages of the number of sunspots have been recorded since 1700, and this time series has served many times as a benchmark in the statistical

literature.^{11,12,13} Here we will apply the method described in the previous section to the sunspot time series in order to compare its performance with that of an optimal ensemble averaging of independently-trained ANNs.⁶

We used the records in the period 1921-1955 as the test set and the remaining ones as training set. Moreover, in order to compare with other ANN studies we considered feedforward networks with 12 input, h hidden and 1 output neurons, and took alternatively $h = 3$ or 4 as described below. Results are appraised in terms of the average relative variance

$$ARV_S = \frac{1}{\sigma_S^2} \mathbb{E}[(t_i - f(\mathbf{x}_i))^2 | (t_i, \mathbf{x}_i) \in S] \quad (7)$$

where S is either the training or test set and σ_S its standard deviation. Here $\mathbf{x}_i = (s_{i-1}, s_{i-2}, \dots, s_{i-12})$ is an input vector, $t_i = s_i$ the associated target output and s_i the mean annual sunspot number for year i .

For the test set above defined, some of the best performances on sunspot numbers reported in the literature correspond to the threshold autoregressive (TAR) model¹¹ ($ARV_{1921-1955} = 0.097$), single ANN approach with weight decay¹³ ($ARV_{1921-1955} = 0.082$), and an optimal ANN ensemble averaging⁶ ($ARV_{1921-1955} = 0.0713$ for feedforward networks). In this latter case the authors trained several 12:4:1 networks with small learning rates (the result quoted corresponds to $\eta = 0.001$) and a large number of training epochs ($N_E = 150,000$). Then, they reconstructed the performance of the aggregate predictor on a random validation set V with 35 points, following its evolution with the number of training epochs of the individual members. Finally, they chose as stopping point of the training processes of all the networks the number of epochs corresponding to the minimum of the aggregate predictor error on V .

In order to compare with Ref. 6, we have considered the same network architecture (12:4:1), learning rate ($\eta = 0.001$) and maximum number of training epochs ($N_E = 150,000$). We generated an aggregate predictor according to the algorithm described in Section 3, using $N_A = 30$. The average over 25 independent runs of the whole procedure produced $ARV_{1921-1955} = 0.0636 \pm 0.0036$, which compares favorably with the corresponding result $ARV_{1921-1955} = 0.0713$ in Ref. 6 (notice that the quoted ARV values are normalized us-

ing the complete record variance $\sigma_D^2 = 1535$ as in Ref. 6, to allow a direct comparison with this work).

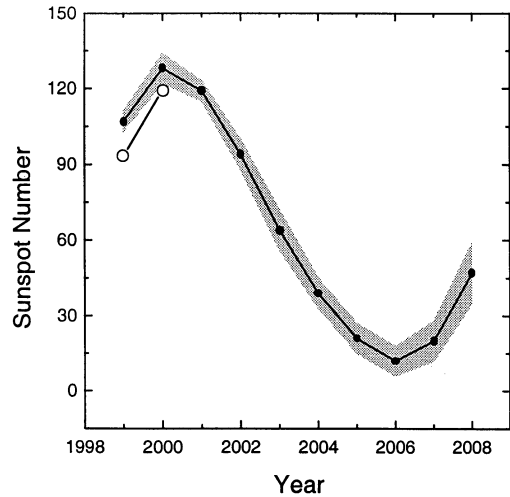


Fig. 1. Predictions for years 1999/2000 and the remaining part of cycle 23 of solar activity. Black dots are the average prediction of the 10 experiments performed and the gray zone indicates the $\pm 2\sigma$ deviation from this average value.

We have also considered the actual task of predicting the remaining part of the current solar cycle 23. Here we took the intervals 1700-1975 and 1987-1998 as data set, and left out cycle 21 corresponding to the period 1976-1986 (which is somehow similar to the current one) for testing. In this preliminary study we used a much smaller maximum number of training epochs ($N_E = 5000$), a larger learning rate ($\eta = 0.01$, and a much smaller $N_A = 4$ (estimated from the performance on cycle 21)). We performed 10 runs of the whole procedure, considering a 12:3:1 network architecture. This experimental setting was chosen in order to compare with results of previous ANN studies of this problem.¹⁴ The experiments produced an average $ARV_{1976-1986} = 0.116$ on the test set, with maximum and minimum values of 0.127 and 0.111, respectively. The average performance of a single network was 0.124. Our forecasts for years 1999/2000 and the remainder of the current solar cycle 23 are shown in Fig. 1, together with the actual observations for years 1999/2000 for comparison. This figure gives the average prediction of the resulting 10 models and the $\pm 2\sigma$ deviation from this average. In spite of the fact that the predicted sunspot numbers for years 1999/2000 are not very

close to the observed values, our result for the maximum of the current solar cycle is among the most accurate ones in the literature (see Ref. 14 for an account of previous works on this problem).

6. Conclusions

We proposed a simple method for balancing diversity and accuracy of ANN ensemble members. At every stage, the algorithm seeks a new member that is at least partially anticorrelated with the previous-stage ensemble estimator. This is achieved by applying a late-stopping method in the training process of individual networks, leading to a controlled level of overtraining of the ensemble members. The algorithm retains the simplicity of independent network training and, moreover, it largely reduces the computational burden compared to other algorithms like NeuralBAG or the method proposed in Ref. 6 (which require saving the intermediate networks during training, since the selection of stopping points for the ensemble members is performed only at the end of all the training processes). Our method is a stepwise construction of the ensemble, where each network is selected at a time and only its parameters have to be saved. We showed, by comparison with other methods proposed in the literature, that this strategy is effective, as exemplified by the results on three standard statistical benchmarks, the Ozone, Boston Housing and Friedman#1 datasets, and on the sunspot time series. We also presented results for the real task of predicting the remainder of the current cycle 23 of solar activity. The results are encouraging and we are presently performing a more extensive check of the algorithm on new databases and different learning conditions.

Acknowledgments

This work was partially funded by ANPCyT of Argentina through PICT 11-03834.

References

1. A. J. C. Sharkey, Ed., *Combining Artificial Neural Nets*, (Springer-Verlag, London, 1999).
2. J. G. Carney and P. Cunningham, "The NeuralBAG algorithm: Optimizing generalization performance in bagged neural networks," *Proc. of the 7th European Symposium on Artificial Neural Networks*, 35-40 (1999); J. Carney and P. Cunningham, "Tuning diversity in bagged ensembles," *International Journal of Neural Systems*, **10**, 267-280 (2000).
3. H. Drucker, R. Schapire and P. Simard, "Improving performance in neural networks using a boosting algorithm," *Advances in Neural Information Processing Systems*, **5**, 42-49 (Morgan Kaufman, 1993).
4. L. Breiman, "Bagging predictors," *Machine Learning*, **24**, 123-140 (1996).
5. Y. Freund and R. Schapire, "A decision-theoretic generalization of on-line learning and an application to boosting," *Proceedings of the 2nd European Conference on Computational Learning Theory*, 23-37 (Springer Verlag, 1995).
6. U. Naftaly, N. Intrator and D. Horn, "Optimal ensemble averaging of neural networks," *Network: Comput. Neural Syst.*, **8**, 283-296 (1997).
7. S. Geman, E. Bienenstock and R. Doursat, "Neural Networks and the Bias/Variance Dilemma," *Neural Computation*, **4**, 1-58 (1992).
8. D. Opitz and J. Shavlik, "Generating accurate and diverse members of a neural network ensemble," *Advances in Neural Information Processing Systems*, **8**, 535-541 (MIT Press, 1996).
9. B. Efron and R. Tibshirani, *An Introduction to the Bootstrap*, (Chapman and Hall, London, 1993).
10. L. Breiman, "Out-of-bag estimation," Technical Report, Statistics Department, University of California at Berkeley (1996).
11. M. B. Priestley, *Spectral Analysis and Time Series*, (Academic Press, New York, 1981).
12. H. Tong, *Threshold models in Non-Linear Time Series Analysis*, Lecture Notes in Statistics 21 (Springer, Berlin, 1983).
13. A. S. Weigend, B. A. Huberman and D. Rumelhart, "Predicting the future: a connectionist approach," *International Journal of Neural Systems*, **1**, 193-209 (1990).
14. P. F. Verdes, M. A. Parodi, P. M. Granitto, H. D. Navone, R. D. Piacentini and H. A. Ceccatto, "Predictions of the maximum amplitude for Solar Cycle 23 and its subsequent behavior using nonlinear methods," *Solar Physics*, **191**, 421-427 (2000).