



## Multi-layer Perceptron and Pruning

C. Voyant<sup>1,2,\*</sup>, C. Paoli<sup>2</sup>, M-L. Nivet<sup>2</sup>, G. Notton<sup>2</sup>, A. Fouilloy<sup>2</sup>, F. Motte<sup>2</sup>

<sup>1</sup>CHD Castelluccio, Radiophysics Unit, 20000 Ajaccio, France

<sup>2</sup>University of Corsica, UMR CNRS 6134 SPE, Campus Grimaldi, BP 52, 20250 Corte, France

### ARTICLE INFO

#### Article history:

Received 18 April 2017

Accepted 23 June 2017

Available online 22 August 2017

#### Keywords:

ANN

Forecasting

Time series

Meteorology

### ABSTRACT

A Multi-Layer Perceptron (MLP) defines a family of artificial neural networks often used in TS modeling and forecasting. Because of its “black box” aspect, many researchers refuse to use it. Moreover, the optimization (often based on the exhaustive approach where “all” configurations are tested) and learning phases of this artificial intelligence tool (often based on the Levenberg-Marquardt algorithm; LMA) are weaknesses of this approach (exhaustively and local minima). These two tasks must be repeated depending on the knowledge of each new problem studied, making the process, long, laborious and not systematically robust. In this short communication, a pruning process is presented. This method allows, during the training phase, to carry out an inputs selecting method activating (or not) inter-nodes connections in order to verify if forecasting is improved. We propose to use iteratively the popular damped least-squares method to activate inputs and neurons. A first pass is applied to 10% of the learning sample to determine weights significantly different from 0 and delete other. Then a classical batch process based on LMA is used with the new MLP. The validation is done using 25 measured meteorological TS and cross-comparing the prediction results of the classical LMA and the 2-stage LMA.

© 2017 Forecast Research Laboratory. All rights reserved.

## 1. Introduction

The primary goal of time series (TS) analysis is forecasting, i.e. using the past to predict the future [1]. This formalism is used in many scientific fields like econometrics, seismology or meteorology [2]. Lots of methods are dedicated to the prediction of discrete phenomena, one of the most popular is the artificial neural network (ANN) [1,3]. From a mathematical point of view, ANN is a function defined as the composition of other functions [4]. Members of the class of such functions are obtained by varying parameters, (as connections or weights) [2]. A Multi-Layer Perceptron (MLP) defines a family of functions often used in TS modeling [1]. In this model, neurons are grouped in layers and only forward connections exist. A typical MLP consists of an input, hidden and output layers, including neurons, weights and a transfer functions. Each neuron (noted  $i$ ) transforms the weighted sum (weight  $w_{ij}$ , bias  $b_i$ ) of inputs ( $x_j$ ) into an output ( $y_i = f(\sum_{j=1}^n x_j w_{ij} + b_i)$ ) using a transfer function ( $f$ ). The goal of this method is to determine weights and bias for a given problem. A complex process is necessary to adapt connections using a suitable training algorithm (often based on the Levenberg-Marquardt algorithm; LMA [5–8]). The training step is dependent on the number of inputs, layers and hidden nodes. The better configuration defines the optimized MLP [1]. This step is the weaknesses of this approach because no consensus or scientific rules exist, often the use of the

\* Corresponding author.

E-mail addresses: [voyant@univ-corse.fr](mailto:voyant@univ-corse.fr) (Cyril Voyant), [christophe.paoli@univ-corse.fr](mailto:christophe.paoli@univ-corse.fr) (Christophe Paoli), [marie-laure.nivet@univ-corse.fr](mailto:marie-laure.nivet@univ-corse.fr) (Marie-Laure Nivet), [gilles.notton@univ-corse.fr](mailto:gilles.notton@univ-corse.fr) (Gilles Notton), [fouilloy\\_a@univ-corse.fr](mailto:fouilloy_a@univ-corse.fr) (Alexis Fouilloy), [motte@univ-corse.fr](mailto:motte@univ-corse.fr) (Fabrice Motte)

exhaustive approach (where "all" configurations are tested) is the only usable and must be repeated for the new studied problem, making the process, long, laborious and not systematically robust. This "black box" aspect leads any researchers to refuse to use it. In this paper, a pruning process allowing to automatically selecting inputs is proposed. This method allows, during the training phase, to carry out a selecting method activating (or not) inter-nodes connections. With this process, the optimization step becomes self-acting and the parsimony principle is kept. Less the MLP is complex more it is efficient [9].

## 2. Materials and Methods

We propose to use iteratively the popular LMA also known as the damped least-squares method to activate the inputs and neurons ( $m$  weights and bias) [9]. A first pass is applied to 10% of the learning sample. For each step, a system of  $m$  non-linear equations with  $m$  unknowns is solved (see equation 1 in the case of 1 hidden layer MLP where  $O$  and  $I$  are the outputs and inputs,  $W^1, B^1$  and  $W^2, B^2$  the weights and bias matrices of the hidden and output layer).

$$O^i = W^2 \cdot (\tanh(W^1 \cdot I^i + B^1)) + B^2 \quad \text{with } 1 < I < m, \quad (1)$$

After this first phase, each weight and bias ( $\omega_{i \in [1, m]}$ ) are represented by probability distributions. A statistical test based on the bootstrap distribution is used to determine if the first moment of each  $\omega_i$  is significantly different from 0 [10]. Before to initiate the second pass, the network is customized and connections related to each  $\omega_i$  non-significantly different from 0 are canceled. Then a classical batch process based on LMA is used with the new MLP. We validated our method with 5 hourly meteorological TS (wind direction  $WD$ , wind speed  $WS$ , Global radiation  $Glo$ , Humidity  $Hum$  and temperature  $Tem$ ), each one measured in 5 French sites (Ajaccio, Bastia, Corte, Marseille and Nice). Note that no pre-treatment are operated and that different non-stationarities and periodicities are present. For all TS and locations, we used 3200 measures for the training and 400 for the cross-comparison of the classical LMA and the 2-stage LMA during the year 2008.

### 2.1. First pass

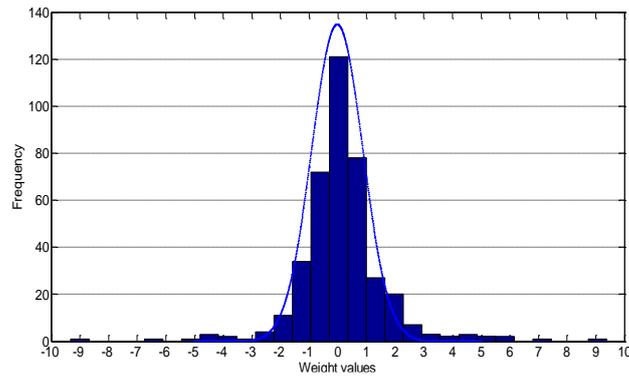
The first stage begins with the generation of  $N$  (10% of the total data used during training randomly chosen) systems of  $m$  nonlinear equations with  $m$  unknowns (MLP constructed with  $m$  weights and bias). The method was chosen for solve this problem is the LMA method and the ad-hoc objective function  $F$  (mean square error between calculations and measures). It is an approximation of the Gauss-newton method, the result of the  $k^{\text{th}}$  iterations ( $\omega_k$ ) corresponding to the local minimum of the function  $F$  is generated by the linear set of equations (1) [6,9,11]:

$$(J(\omega_k)^T \cdot J(\omega_k) + \lambda_k I) \cdot d_k = -J(\omega_k)^T \cdot F(\omega_k) \quad (2)$$

$J$  denotes the Jacobian matrix of  $F$  and the scalar  $\lambda_k$  controls both the magnitude and variation  $d_k$  ( $d_k = \Delta\omega = \omega_{k+1} - \omega_k$ ). After the  $N$  solving, all the weights and bias are represented by a distribution which will be used during the second stage of the methodology. Note that these distributions are not all normal (according to the Jarque-Bera test).

### 2.2. Second pass

Before to initiate the second pass, the network is customized and connections related to each  $\omega_i$  non-significantly different from 0 are cancelled. An example of the distribution related to one weight is available in Figure 1.

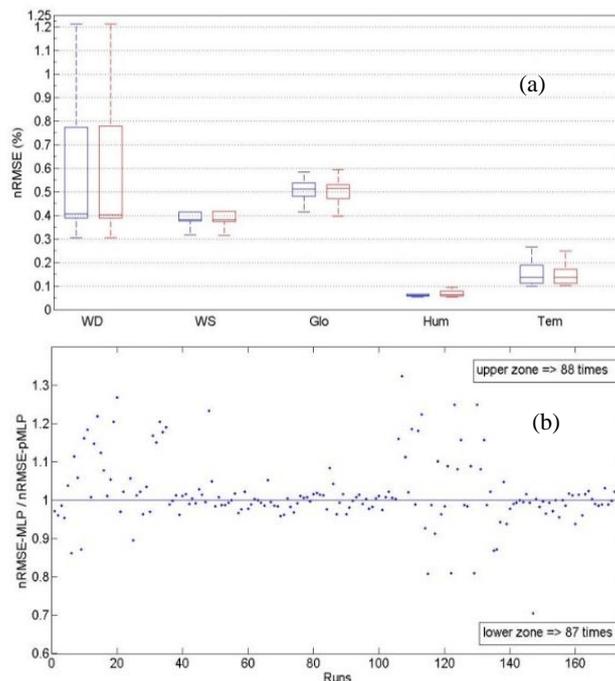


**Figure 1.** Example of an MLP weight distribution

To perform the second pass, we use confidence intervals from the bootstrap distributions (4000 samples) of the weights and bias parameters [10]. The rules of selection (directly linked to the  $\alpha$  value of significance model) is based on the product of the two endpoints  $t_1$  and  $t_2$  defined respectively by the  $(\alpha/2)^{\text{th}}$  and  $(1 - \alpha/2)^{\text{th}}$  percentiles of the distribution. If  $t_1 \cdot t_2 < 0$  the weight (or bias) is considered as non-significantly different from 0, else it is considered different from 0. The connection of the MLP corresponding to the first case ( $t_1 \cdot t_2 < 0$ ) are cancelled, other are kept. The pruned MLP (noted pMLP) is then trained with the classical LMA.

### 3. Results

In the Figure 2a is represented the box plot of the normalized root mean square error (nRMSE) distribution concerning the five meteorological parameters and the five studied cities. For each case, seven runs are operated, so 175 manipulations are performed with the pruned methodology described above (at left in Figure 1 and noted pMLP) and the standard approach (at right in the Figure 1 and noted MLP). The chosen architecture is the same for all cases: 7 inputs representing the seven first lags of the meteorological parameter tested and 2 hidden nodes (only one hidden layer). In this figure, we see that the first, the second and the third quartiles are equivalent; thereby we understand that a lot of connections and weights are superfluous. The results of all the simulations are represented in the Figure 2b. Only the points positioned in the upper zone are related to “pMLP is better than MLP” cases are plotted. It appears the points are closer to the  $y=1$  curve in the top area rather than in the bottom area, but this observation seems insignificant.



**Figure 2.** (a). nRMSE distribution comparison related to pMLP (at left) and MLP (at right), (b). the ratio of the nRMSE generated by MLP and pMLP

The table 1 and 2 expose the results for all locations and parameters of the MLP and pMLP approach, the minimum of the nRMSE and nMAE (normalized mean absolute error) [1] through the seven runs is exposed in the first and the average in the second.

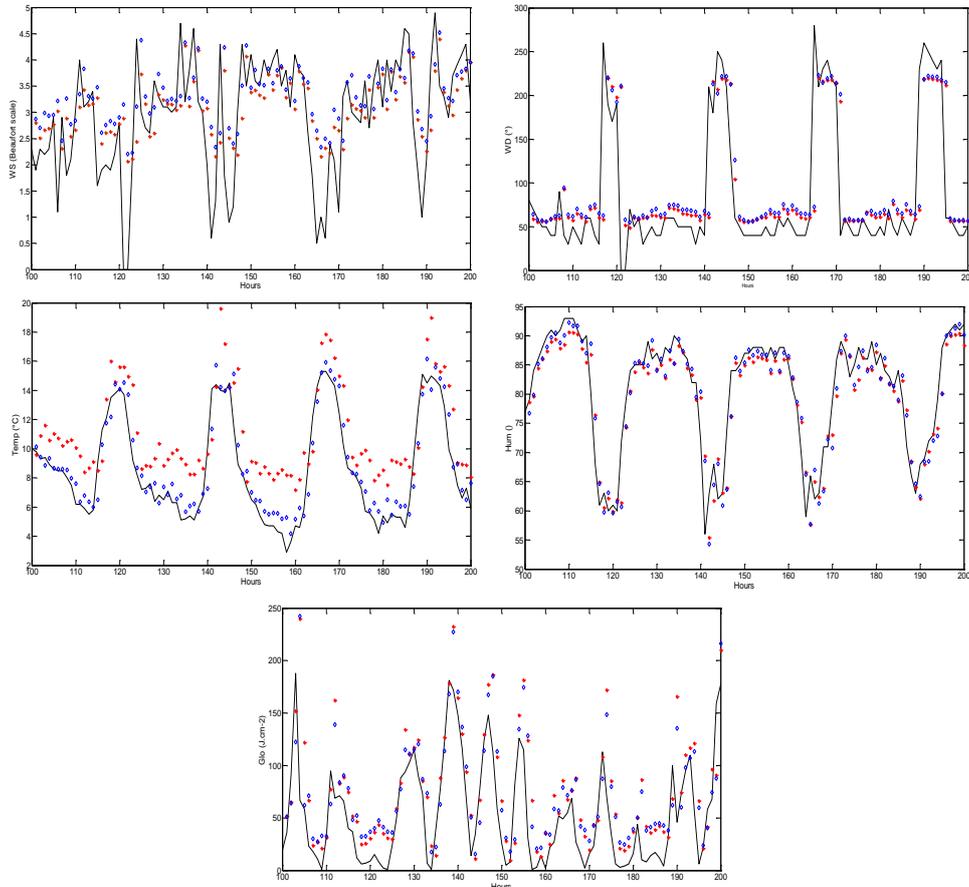
**Table 1.** nRMSE and nMAE minima for all location and all parameters, in bold the better results between MLP and pMLP

Data	City	MLP		pMLP		
		nRMSE	nMAE	Ratio pruning	nRMSE	nMAE
WD	Aja	0.765	0.463	0.27	<b>0.762</b>	<b>0.461</b>
	Bas	0.393	0.257	0.27	<b>0.388</b>	<b>0.252</b>
	Cor	<b>1.191</b>	0.941	0.21	1.194	<b>0.931</b>
	Mar	<b>0.387</b>	<b>0.256</b>	0.20	0.388	0.257
	Nic	<b>0.304</b>	<b>0.199</b>	0.20	0.305	0.211
WS	Aja	<b>0.399</b>	<b>0.302</b>	0.11	0.410	0.308
	Bas	0.374	<b>0.277</b>	0.18	<b>0.370</b>	0.279
	Cor	<b>1.148</b>	0.910	0.22	1.183	<b>0.902</b>
	Mar	<b>0.377</b>	0.264	0.21	<b>0.377</b>	<b>0.263</b>
	Nic	0.318	0.206	0.23	<b>0.314</b>	<b>0.202</b>
Glo	Aja	0.525	0.413	0.17	<b>0.472</b>	<b>0.372</b>
	Bas	<b>0.439</b>	<b>0.323</b>	0.21	0.456	0.370
	Cor	<b>0.298</b>	<b>0.214</b>	0.21	0.323	0.250
	Mar	0.416	0.346	0.21	<b>0.378</b>	<b>0.302</b>
	Nic	<b>0.455</b>	<b>0.380</b>	0.20	0.465	0.386
Hum	Aja	<b>0.064</b>	0.049	0.22	<b>0.064</b>	<b>0.048</b>
	Bas	<b>0.061</b>	<b>0.044</b>	0.20	<b>0.061</b>	0.045
	Cor	<b>0.055</b>	<b>0.036</b>	0.23	0.056	0.037
	Mar	<b>0.053</b>	<b>0.036</b>	0.23	<b>0.053</b>	<b>0.036</b>
	Nic	<b>0.083</b>	<b>0.059</b>	0.19	<b>0.083</b>	<b>0.059</b>
Tem	Aja	<b>0.099</b>	<b>0.077</b>	0.29	0.101	<b>0.077</b>
	Bas	0.113	0.080	0.25	<b>0.111</b>	<b>0.079</b>
	Cor	0.224	0.158	0.29	<b>0.207</b>	<b>0.133</b>
	Mar	0.111	0.079	0.28	<b>0.109</b>	<b>0.073</b>
	Nic	0.147	<b>0.108</b>	0.21	<b>0.146</b>	0.111

**Table 2.** nRMSE and nMAE mean for all location and all parameters, in bold the better results between MLP and pMLP

Data	City	MLP		pMLP		
		nRMSE	nMAE	Ratio pruning	nRMSE	nMAE
WD	Aja	<b>0.772</b>	<b>0.478</b>	0.27	0.774	0.484
	Bas	0.418	0.287	0.27	<b>0.401</b>	<b>0.270</b>
	Cor	<b>1.197</b>	0.972	0.21	1.202	<b>0.958</b>
	Mar	<b>0.391</b>	<b>0.262</b>	0.20	0.394	0.269
	Nic	<b>0.308</b>	<b>0.212</b>	0.20	0.310	0.218
WS	Aja	<b>0.409</b>	<b>0.309</b>	0.11	0.414	0.312
	Bas	0.378	<b>0.280</b>	0.18	<b>0.376</b>	0.281
	Cor	1.203	0.960	0.22	<b>1.196</b>	<b>0.915</b>
	Mar	<b>0.379</b>	<b>0.266</b>	0.21	0.380	<b>0.266</b>
	Nic	0.323	0.212	0.23	<b>0.322</b>	<b>0.210</b>
Glo	Aja	0.574	0.465	0.17	<b>0.562</b>	<b>0.400</b>
	Bas	0.551	<b>0.453</b>	0.21	<b>0.386</b>	0.467
	Cor	<b>0.389</b>	<b>0.304</b>	0.21	0.464	0.311
	Mar	<b>0.481</b>	0.410	0.21	0.514	<b>0.393</b>
	Nic	0.489	<b>0.413</b>	0.20	<b>0.400</b>	0.434
Hum	Aja	<b>0.069</b>	<b>0.054</b>	0.22	0.073	0.056
	Bas	<b>0.061</b>	<b>0.045</b>	0.20	0.062	0.046
	Cor	<b>0.057</b>	<b>0.039</b>	0.23	0.058	<b>0.039</b>
	Mar	<b>0.054</b>	<b>0.037</b>	0.23	<b>0.054</b>	<b>0.037</b>
	Nic	0.085	0.061	0.19	<b>0.084</b>	<b>0.060</b>
Tem	Aja	<b>0.109</b>	<b>0.087</b>	0.29	0.112	0.089
	Bas	0.139	0.111	0.25	<b>0.128</b>	<b>0.102</b>
	Cor	0.259	0.205	0.29	<b>0.236</b>	<b>0.176</b>
	Mar	<b>0.112</b>	<b>0.081</b>	0.28	0.126	0.090
	Nic	0.179	0.138	0.21	<b>0.159</b>	<b>0.121</b>

If the mean of the error metrics is interesting to compare the global trend, the minimum values allow to determine the best learning, and therefore the best network. pMLP is very slightly better than MLP, in the minima case, 60% of cases the nRMSE and the nMAE are the lowest. Note that the pruning concerns about 20% of weights and bias. In Figure 3 is available the predictions related to the five time series compares to the measurements.



**Figure 3.** Profile during 100 days concerning measurements (lines) MLP (cross) and pMLP (circles)

In this figure, pMLP gives better visual results for temperature, however concerning the four other time series the gain is not significant.

## 4. Conclusions

The 2-pass approach improves slightly the forecasting quality. In average, 20% of the connections are removed with this approach. According to the parsimony principle, these simplifications increase the generalization capacity and should allow building a robust predictor [3].

This first study was done in a quasi-optimized case (7 inputs and 2 hidden neurons) precedes a more general one, where a standard MLP (more than 15 inputs and 15 hidden nodes) will be studied. Indeed, we have shown that the pruning method presented here is able to simplify the network while the performance is roughly equivalent. Applying the 2-pass approach in a 15x15 MLP should allow to optimize it without applying the exhaustive test where all the architectures are trying out. Moreover, for users, it is a totally transparent methodology, suitable for all TS and faster than classical optimization process. According to the conclusion of this study, it possible that results based on the 2-stage approach may be better than the classical approach based on the "1-stage" LM algorithm.

## References

- [1] C. Voyant, M. Muselli, C. Paoli, M.L. Nivet, Numerical weather prediction (NWP) and hybrid ARMA/ANN model to predict global radiation, *Energy*. 39 (2012) 341–355. doi:10.1016/j.energy.2012.01.006.
- [2] C. Voyant, G. Notton, C. Paoli, M.L. Nivet, M. Muselli, K. Dahmani, Numerical weather prediction or stochastic modeling: an objective criterion of choice for the global radiation forecasting, *Int. J. Energy Technol. Policy*. (2014).

- [3] A. Mellit, S.A. Kalogirou, L. Hontoria, S. Shaari, Artificial intelligence techniques for sizing photovoltaic systems: A review, *Renew. Sustain. Energy Rev.* 13 (2009) 406–419. doi:10.1016/j.rser.2008.01.006.
- [4] G. Cybenko, Correction: Approximation by Superpositions of a Sigmoidal Function, *Math. Control. Signals, Syst.* 2 (1989) 303–314. doi:10.1007/BF02134016.
- [5] J. Fan, J. Pan, A note on the Levenberg-Marquardt parameter, *Appl. Math. Comput.* 207 (2009) 351–359. doi:10.1016/j.amc.2008.10.056.
- [6] C.K. Yoo, S.W. Sung, I.B. Lee, Generalized damped least squares algorithm, *Comput. Chem. Eng.* 27 (2003) 423–431. doi:10.1016/S0098-1354(02)00219-3.
- [7] C. Voyant, W. Tamas, C. Paoli, A. Balu, M. Muselli, M.-L. Nivet, G. Notton, Time series modeling with pruned multi-layer perceptron and 2-stage damped least-squares method, in: *J. Phys. Conf. Ser.* 490, Institute of Physics Publishing, Prague, 2014: pp. 1–4. doi:10.1088/1742-6596/490/1/012040.
- [8] C. Voyant, W. Tamas, M.-L. Nivet, G. Notton, C. Paoli, A. Balu, M. Muselli, Meteorological time series forecasting with pruned multi-layer perceptron and two-stage Levenberg-Marquardt method, *Int. J. Model. Identif. Control.* 23 (2015) 287–294. doi:10.1504/IJMIC.2015.069952.
- [9] H. Brusset, D. Depeyre, J.-P. Petit, F. Haffner, On the convergence of standard and damped least squares methods, *J. Comput. Phys.* 22 (1976) 534–542. doi:10.1016/0021-9991(76)90048-6.
- [10] J.-P. Kreiss, E. Paparoditis, Bootstrap methods for dependent data: A review, *J. Korean Stat. Soc.* 40 (2011) 357–378. doi:10.1016/j.jkss.2011.08.009.
- [11] F.M. Dias, A. Antunes, J. Vieira, A. Mota, A sliding window solution for the on-line implementation of the Levenberg-Marquardt algorithm, *Eng. Appl. Artif. Intell.* 19 (2006) 1–7. doi:10.1016/j.engappai.2005.03.005.