A New Decision Strategy in Multi-Objective Training of Artificial Neural Networks

Talles H. Medeiros and Ricardo H. C. Takahashi and Antônio P. Braga *

1- Federal University of Minas Gerais - Department of Electronic Engineering
Belo Horizonte, MG, Brazil, email: talles, apbraga@cpdee.ufmg.br)

2- Federal University of Minas Gerais - Department of Mathematics
Belo Horizonte - MG, Brazil (email: taka@mat.ufmg.br)

Resumo. In this work it is presented a new proposal to select a model in the multi-objective training method of the Artificial Neural Network (NN). In order to do this, information from the residue of the Pareto optimal solution is used. The principle to decide for minimum autocorrelation of the data is a criteria that guarantees the extraction of the current information in the noisy data. The experiments show the performance of the proposed DM for variations of the supervised learning problems.

1 Introduction

Although the single objective, error minimization, approach to supervised learning may, eventually, result on good generalization responses, it is well accepted that training should also include model complexity minimization. Despite there is no clear definition of complexity in the literature, it is usually associated with the total number of network parameters. So, much of the last decades efforts in the area were concerned on minimizing error and network size by using pruning or constructive techniques [1]. Smaller size models were selected according to their responses on validation and test sets. Model selection with alternative approaches, such as cross-validation, bagging and regularization [1] did not explicitly refer to structural complexity, since any large enough network could meet the selection criteria. In this approach, emphasis is given to the network response instead to its actual size, so a large network may effectively behave like a smaller one.

Smoothing and sampling restrictions imposed during training have the effect of restricting the search to a limited region of the space of solutions. Although this is not accomplished explicitly, in practice, it has also the effect of reducing the network behaviour to those solutions that are within the restricted region. From this perspective, it is expected to exist an objective function that is able to control network effective complexity without restricting its size. The alternative for this is to use also the norm $\|w\|$ of network weight vectors to control network response. The effect of imposing an upper bound into the values that can be assumed by the network norm is to restrict its effective complexity, since a network with a larger norm may assume also all the solutions of smaller norm networks. So, the larger the norm, the larger the number of network solutions that can be assumed.

*(phone: +55 31 3499-4869; fax: +55 31 3499 4850. http://www.litc.cpdee.ufmg.br)

The relationship between norm and margin of large margin classifiers is one of the basic premisses of SVMs formulation [1]. Its relationship with network generalization has also been shown in another context [2]. This supports the idea that the trade-off between bias and variance [3] could be controlled by an appropriate balance between norm $\|\mathrm{w}\|$ and error e^2 of the training set, since simultaneous minimization of both objectives in the region of interest is not possible. Minimum error corresponds to large variance/minimum bias and, similarly, minimum norm corresponds to large bias/minimum variance. Therefore, the pair of objectives (norm,error) and (bias,variance) are conflicting near the objective functions minima, what demands a proper treatment that is able to trade-off them. The region of effective solutions in the space of objectives, that is called the Pareto set, can be reached by using Multi-objective optimization approaches [4].

Based on these principles, a multi-objective optimization algorithm for Multi-Layer Perceptron (MLPs) has been proposed in [5]. This algorithm employs a constrained optimization approach to restrict the solutions to the Pareto set efficient solutions, defined by the two objective functions $\|\mathrm{w}\|$ and e^2. Every network solution corresponds to a pair (e^2, $\|\mathrm{w}\|$) on the space of objectives (see Figure 1. The Pareto set is found in the boundary between the image set of the vector function ($e^2(w)$, $\|\mathrm{w}\|$) and the set of points outside this image. The Pareto set contains the *efficient solutions*, that cannot be further minimized in both objectives. Every pair (e_i^2, $\|\mathrm{w_i}\|$) on the Pareto set corresponds to the solution of minimum $\|\mathrm{w_i}\|$ for a given e_i^2, or to a solution of minimum error e_i^2 for a given norm $\|\mathrm{w_i}\|$. Solutions that are above the Pareto set can be confronted with other ones that have both $\|\mathrm{w}\|$ and e^2 smaller.

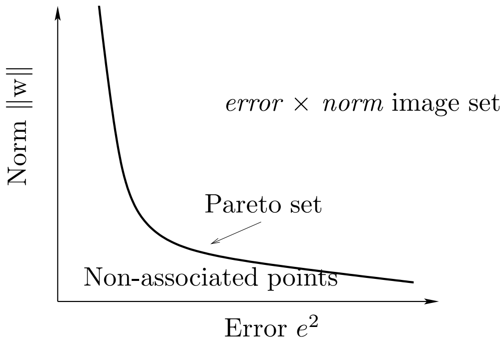

Figura 1: Each point in the *error* × *norm* image set corresponds to the pair ($\|\mathrm{w}\|$,e^2) that is associated to at least one MLP. There are no MLPs associated to pairs ($\|\mathrm{w}\|$,e^2) in the region of the non-associated points.

Once all the Pareto set solutions are generated, one of them is selected, according to a pre-established decision making criteria. In the original multi-objective learning proposed in [5], the Pareto set is sampled via successive constrained optimizations, and the *decider* picks up the solution with the smallest validation error. Although it has been shown that the validation error curve has a minimum within the Pareto set range [5], this original strategy has the drawback of relying on the availability of a validation data set. In this paper, a new decider is proposed, that is based on the idea of minimizing the autocorrelation of residues within the Pareto set solutions. This eliminates the need for a validation data set, allowing all the data to be used on training.

2 The MOBJ Algorithm

A multi-objective optimization algorithm [1] for MLPs has been proposed in [5].

$$\begin{array}{l} \min\limits_{\mathrm{w}} \; \frac{1}{N} \sum\limits_{j=1}^{N} (\mathrm{y_j} - f(\mathrm{x}_j; \mathrm{w}))^2 \\ \text{subject to}: \; \|\mathrm{w}\| \leq \sigma \end{array} \tag{1}$$

On the above, w is the weight vector, N is the length of the training set, $\mathrm{x_j}$ and $\mathrm{y_j}$ are, respectively, j-the input and output example of training set. For calculating a number ζ of Pareto-optimal solution, it is necessary to define this number and variation σ. The chosen solution is the best trade-off between the norm and the error is performed after determining some Pareto-optimal solution and is named *Decisor Making* (DM).

3 The Decision Making

The DM is expected to choose the solution that best fits the underlying function $f_g(\mathrm{x})$. Consider the supervised learning of a MLP, with the training examples given by sample points x_i in a domain D and corresponding sample values y_i:

$$\{(\mathrm{x}_i, \mathrm{y}_i) | \mathrm{y_i} = f_g(\mathrm{x}) + \xi_i\}_{i=1}^{N} \tag{2}$$

where ξ is i.i.d. zero mean random error (noise), x is a multidimensional input and y is a scaled output. The estimation made is based on a finite number (N) of the training data. The training data is independent and identically distributed (i.i.d.) according to some (unknown) joint probability density function.

Let w be a weight vector, that is obtained via a learning procedure, and $f(\mathrm{x}, \mathrm{w})$ be the result obtained associated to the solution w. The DM can be stated as: *From a given set of Pareto-optimal solution, to find the minimized solution in this set, the function is defined as:*

$$U = \|f_g(\mathrm{x}) - f(\mathrm{x}, \mathrm{w}_i)\|^2. \tag{3}$$

The final solution generates a residual error with variance σ^2 [6]. Therefore, equivalence between $f_g(\mathrm{x})$ and $f(\mathrm{x}; \mathrm{w})$ makes that the residue of each solution after the training, corresponds to noise ξ for best solution and its variance is determined by $\frac{1}{N}\sum_{i=1}^{N} (\xi_i)^2 = \sigma^2$.

[1] See [4] for reference of multi-objective optimization.

4 The Proposed Decisor Making

The decision method that is proposed here uses the residue left from the difference between the training data and the outputs of the resulting MLP. The optimal solution, under the proposed criteria, will bring closer the target function $f_g(\mathrm{x})$ that underlies the data $\mathrm{y} = f_g(\mathrm{x}) + \xi$ like the remaining residue ξ becomes not correlated The DM is based on minimizing the autocorrelation residue, over the set of the Pareto-optimal solution.

The decision process chooses the MLP with the residue that is more similar to a random process. Such process has small correlation ($\mathrm{R_{xx}} \to 0$). The correlation measured [7] of the data set can be used to detect non-randomness. This supports the idea of minimal autocorrelation as an indicator of the MLPs with best generalization. The autocorrelation measured from a pair of variables of the same stationary process taken from the same process X is given by:

$$R_{XX}(t, t+\tau) = R_{XX}(\tau) = E[X(t), X(t+\tau)], \tag{4}$$

where $E[.]$ is a expectation of the random variable. Taking the variable τ as an offset between the two samples of the same stationary process. The decision rule, by minimal autocorrelation, is given by:

$$\mathrm{w}^* = \underset{\mathrm{w} \in \mathcal{W}^*}{\arg} \quad \min \ R_{\mathrm{xx}}, \tag{5}$$

with R_{xx} given by:

$$R_{\mathrm{xx}} = E[(d_i + \xi_i) - f(\mathrm{x}_i;\ \mathrm{w}),\ \ (d_i + \xi_i) - f(\mathrm{x}_i;\ \mathrm{w}) + \tau]. \tag{6}$$

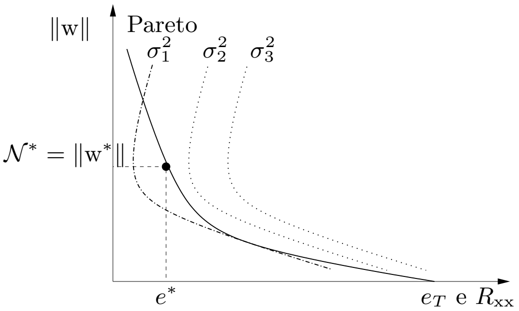

Figura 2: Pareto set and autocorrelation curves for different noises.

5 Results

The MLP is designed with more hidden neurons than the minimal number that would be needed, in order to show over-fitting effects smoothed by the multi-objective training with autocorrelation DM. The functions f_1 and f_2 with Gaussian noise have been used as targeted functions, with input values scattered in different intervals.

$$f_1(\mathrm{x}) = \sin(\mathrm{x}) \qquad f_2(\mathrm{x}) = 4.26(e^{-\mathrm{x}} - 4e^{-2\mathrm{x}} + 3e^{-3\mathrm{x}}) \tag{7}$$

A MLP with topology 1-10-1 (31 parameters) has been designed by multi-objective training with minimum autocorrelation and minimum validation error DMs. The activation function is sigmoidal (hidden neurons) and linear (output neuron). The MLPs have been trained with 50 noisy patterns. The MOBJ method generated 20 Pareto-optimal solutions. The Table 1 and 2 show results for numerical experiments with different sampling and different noises. The Figure 5 show the final solution by the minimum correlation DM for mapping the targeted function (f_1) with variance noise $\sigma^2 = 0.1$.

Tabela 1: Approximation with noise patterns $\sigma^2 = 0.20$ and distinct sampling.

Points	$\|\mathrm{w}\|$	R_{xx}	Error (MSE)
50	2.9976	8.7250	0.0980
40	5.9995	7.4902	0.0539
30	3.9968	5.6357	0.1131
20	2.9996	6.3902	0.1200
10	7.0015	3.4630	0.1792

Tabela 2: Results for numerical experiments with different noises.

Noise	$\sum e^2$	$\|\mathrm{w}\|$
0.10	0.0222 ± 0.0299	4.9904 ± 0.7968
0.20	0.0511 ± 0.0275	4.6457 ± 0.2462
0.25	0.0640 ± 0.0155	4.6954 ± 0.2925
0.30	0.0859 ± 0.0213	4.9723 ± 1.1640

6 Conclusions

The objective of this paper was to present results from a decision procedure based on the minimization of the autocorrelation residue over the Pareto-set of *error* × *norm* multi-objective training problem. The proposed DM does not need any previous information about the problem. Compared to the DM by minimum validation error, the new DM does not depend on any validation data, which allows a consistent training procedure with less training data. The proposed DM seems to be robust against high noise levels in the data and guarantee

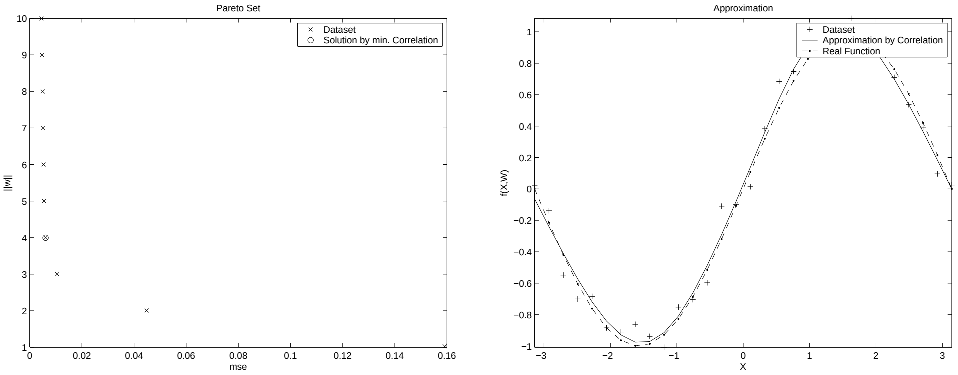

Figura 3: Pareto Set and Best Solution (Best Approximation).

high capability generalization of neural models. Another important factor to be highlighted in the experiments is the DMs capacity to remain, regardless of the complexity, or either, the final solution which did not change very much in the experiments with training sets of different noise levels.

Referências

[1] S. Haykin. *Neural networks: A comprehensive foundation.* MacMillan, New York, 1994.

[2] P. L. Bartlett. For valid generalization, the size of the weights is more important than the size of the network. In *Advances in Neural information Processing Systems*, pages 134–140. MIT Press, 1997.

[3] S. Geman, E. Bienenstock, and R. Doursat. Neural networks and bias/variance dilemma. *Neural Computing*, 4(1):1–58, 1992.

[4] Y. Sawaragi, H. Nakayama, and T. Tanino. *Theory of Multiobjective Optimization.* Academic Press, 1985.

[5] R. A. Teixeira, A. P. Braga, R. H. C. Takahashi, and R. R. Saldanha. Improving generalization of MLPs with multi-objective optimization. *Neurocomputing*, 35(1-4):189–194, 2000.

[6] R. A. Teixeira, A. P. Braga, R. H. C. Takahashi, and R. R. Saldanha. Decisior implementation in neural model selection by multi-objective optimization. In *7th (SBRN), 11-14 November 2002, Recife, Brazil*, page 234. IEEE Computer Society, 2002.

[7] G. E. P. Box and G. Jenkins. *Time Series Analysis, Forecasting and Control.* Holden-Day, Incorporated, 1990.