ESANN'2002 proceedings - European Symposium on Artificial Neural Networks
Bruges (Belgium), 24-26 April 2002, d-side publi., ISBN 2-930307-02-1, pp. 289-294

Evaluation of gradient descent learning algorithms with adaptive and local learning rate for recognising hand-written numerals

M. Giudici (*), F. Queirolo (*) and M. Valle (**)

(**) Department of Biophysical and Electronic Engineering
University of Genova, Via all'Opera Pia 11/a, 16145 Genova, Italy
Ph.: +39 010 353 2775, fax: +39 010 353 2777, email: valle@dibe.unige.it

(*) Department of Production Engineering
University of Genova, Via Cadorna 2, 17100 Savona, Italy
Ph.: +39 019 264555, fax: +39 019 264558

Abstract

Gradient descent learning algorithms, namely Back Propagation (BP), can significantly increase the classification performance of Multi Layer Perceptrons adopting a local and adaptive learning rate management approach. In this paper, we present the comparison of the performance on hand-written characters classification of two BP algorithms, implementing fixed and adaptive learning rate. The results show that the validation error and average number of learning iterations are lower for the adaptive learning rate BP algorithm.

1. Introduction

Neural Networks (NNs) are an efficient computational paradigm for solving many real world problems. At present there is a growing interest in applications like Optical Characters Recognition (OCR), remote-sensing images classification, industrial quality control analysis and many others in which NNs can be effectively employed [1]. Most widely used and reliable learning algorithms are gradient descent ones, namely Back Propagation (BP).

In a previous paper, we have presented an algorithm for the local and adaptive management of the learning rate parameter [2]. Now the aim of this paper is to fully validate the cited algorithm in a meaningful application (i.e. recognition of handwritten characters/numerals) and to demonstrate that the algorithm can significantly increase the performances of the usual BP algorithm. In this paper we present the comparison between the performances of the Fixed Learning Rate (F-LR) and Adaptive and local Learning Rate (A-LR) BP algorithm with Multi Layer Perceptrons (MLPs) networks. The simulation results show that the A-LR algorithm classification and generalization performances are significantly better than F-LR ones.

This paper is organized as follows: in Section 2 the A-LR BP algorithm is introduced, while in Section 3 the characters database and the experimental methodology are presented along with the simulation results. The conclusions are drawn in Section 4.

ESANN'2002 proceedings - European Symposium on Artificial Neural Networks
Bruges (Belgium), 24-26 April 2002, d-side publi., ISBN 2-930307-02-1, pp. 289-294

2. The adaptive and local learning rate BP algorithm

The NN topology we refer in this paper is the two-layer MLP. MLPs can be trained by using the BP operated by pattern, where the training set consists of exemplary input patterns and desired output patterns (i.e. targets). The weights are updated until the target output is generated for the corresponding network input [1].
When considering supervised "gradient descent" based learning algorithms, the learning task is accomplished by minimising, with respect to the adjustable parameters **w** (i.e. weights set) the error index $\varepsilon(\mathbf{w})$; the weight update learning rule is:

$$\Delta w_{i,j} = -\eta \frac{\partial \varepsilon(\mathbf{w})}{\partial w_{i,j}}$$

where η is the Learning Rate (LR) and $w_{i,j}$ is the generic weight connecting neuron i and j, taking in account the generic synapse connecting the i^{th} neuron of the l^{th} layer and the j^{th} neuron of the $(l-1)^{th}$ layer. By applying the chain rule, the BP algorithm derives mathematically the value of $\partial\varepsilon(\mathbf{w})/\partial w_{i,j}$.
To accelerate the learning process, we can implement an adaptive and local learning rate management strategy [2]: each synapse w_{ij} has its own local learning rate η_{ij} and the value of each learning rate is changed adaptively following the behavior of the local gradient error function ($\partial\varepsilon/\partial w_{ij}$). More precisely, η_{ij} is increased when during at least two successive iterations, the signs of the term $\partial\varepsilon/\partial w_{ij}$ are equal, and it is decreased when the signs of the term $\partial\varepsilon/\partial w_{ij}$ are opposite.
So the local learning rate update rule can be formulated as follows:

$$\eta_{i,j}{}^{l}(t+1) = \begin{cases} \eta_{i,j}{}^{l}(t)\left[\dfrac{\eta^{\max}}{\eta_{i,j}{}^{l}(t)}\right]^{\gamma} & when\ S^{l}{}_{ij}(t) = S^{l}{}_{ij}(t-1) \\ \eta_{i,j}{}^{l}(t)\left[\dfrac{\eta^{\min}}{\eta_{i,j}{}^{l}(t)}\right]^{\gamma} & when\ S^{l}{}_{ij}(t) \neq S^{l}{}_{ij}(t-1) \end{cases}$$

$$where\ \ S^{l}{}_{ij}(t) = sign(\partial\varepsilon / \partial w_{ij}{}^{l}) = \begin{cases} +1 & if\ \partial\varepsilon / \partial w_{ij}{}^{l} > 0 \\ -1 & if\ \partial\varepsilon / \partial w_{ij}{}^{l} < 0 \end{cases}$$

where, $S^{l}_{ij}(t)$ is the sign of the gradient component $\partial\varepsilon(t)/\partial w_{ij}{}^{l}$, $\eta^{l}_{ij}(t)$ is the learning rate value at the t^{th} iteration; η^{max} and η^{min} are respectively the maximum and minimum values of the learning rate and γ is the learning rate adaptation coefficient ($\gamma \in [0,1]$).

ESANN'2002 proceedings - European Symposium on Artificial Neural Networks
Bruges (Belgium), 24-26 April 2002, d-side publi., ISBN 2-930307-02-1, pp. 289-294

3. Recognition experiments

3.1 Database

The NIST 19 database [3] was used for the experiments: it contains 402953 samples of segmented handwritten digits, each one represented by a 128×128 binary pixels matrix. Each character matrix is then normalized to 32x24 pixels size. Then, a feature extraction algorithm is applied [4]. The character matrix is partitioned into 16 sub-matrices of size 9×7 pixels, partially overlapped by one or two columns and/or pixel rows. Let us consider now a single sub-matrix. Seven Pattern Recognition (PR) operators are applied on each sub-matrix to obtain a set of seven features. Since each character-matrix is partitioned into 16 sub-matrices, the overall number of features for each input character is then 112.
Out of NIST database, we consider 8000 numerals (800 per class) partitioned into three sub-sets: the training set (4000 numerals), the validation set (2000 numerals) and the test set (2000 numerals). The parameter γ was set at 0.5.

3.2 Methodology

The learning phase has been stopped when the error computed on the validation set (i.e. Validation Error – VE) started to get worse (i.e. at the overfitting point), according to [1]. The test set was used after the training phase to compute the generalisation error (GE) at the overfitting point. The average number of epochs at the overfitting point is referred in the following as Average Training Iterations (ATI). For the GE evaluation we fix the weights configuration at the overfitting point.
Please note that each data point in the following charts has been computed as the average on ten simulation runs with different, random initial weights configurations.

3.3 Simulation results

To evaluate the performance of F-LR and A-LR BP algorithm, we analysed the results coming from experiments on a large number of network topologies (i.e. with different numbers of Hidden Neurons – HN). For each network topology, we determined the VE, the standard deviation of VE, the ATI and the standard deviation of ATI. We also evaluated the GE standard deviation.

The F-LR and A-LR values that were used in the experiments are:

i) F-LR: $\eta = 0.5,\ 0.1,\ 0.01,\ 0.001$;
ii) A-LR: $\eta_{ij} \in [0.5, 0.1], [0.1, 0.01], [0.01, 0.001], [0.1, 10^{-6}]$.

From previous experiments, we identified the network parameters which give the best results in terms of:

i) VE;
ii) ATI;
iii) Standard deviation of VE;
iv) Standard deviation of ATI.

ESANN'2002 proceedings - European Symposium on Artificial Neural Networks
Bruges (Belgium), 24-26 April 2002, d-side publi., ISBN 2-930307-02-1, pp. 289-294

They correspond to:

i) F-LR: $\eta = 0.01$;

ii) A-LR: $\eta_{ij} \in [0.1, 0.01]$.

Figure 1 shows the VE and the ATI at different HN values with the F-LR algorithm (i.e. $\eta = 0.01$); Figure 2 presents the results for the same experiments, but with the A-LR algorithm (i.e. $\eta_{ij} \in [0.1, 0.01]$).

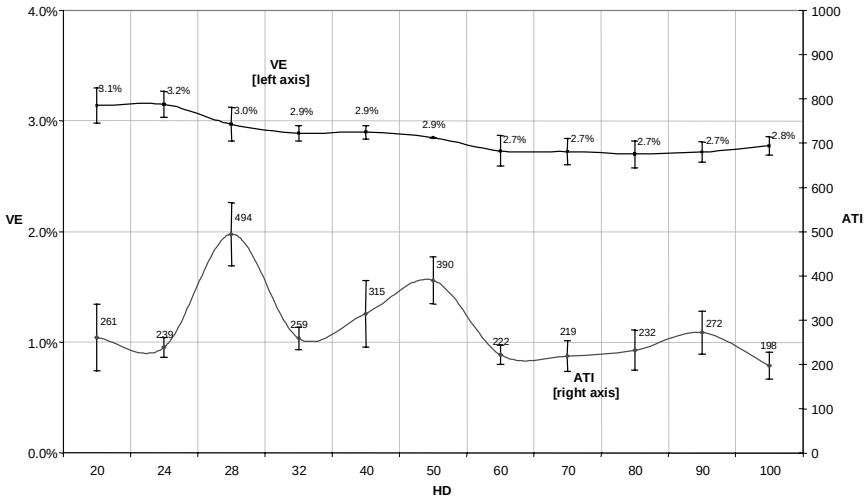

Figure 1: VE (left axis) and ATI (right axis) versus the number of HN (F-LR with $\eta = 0.01$). The vertical bars at each data point indicate the standard deviation.

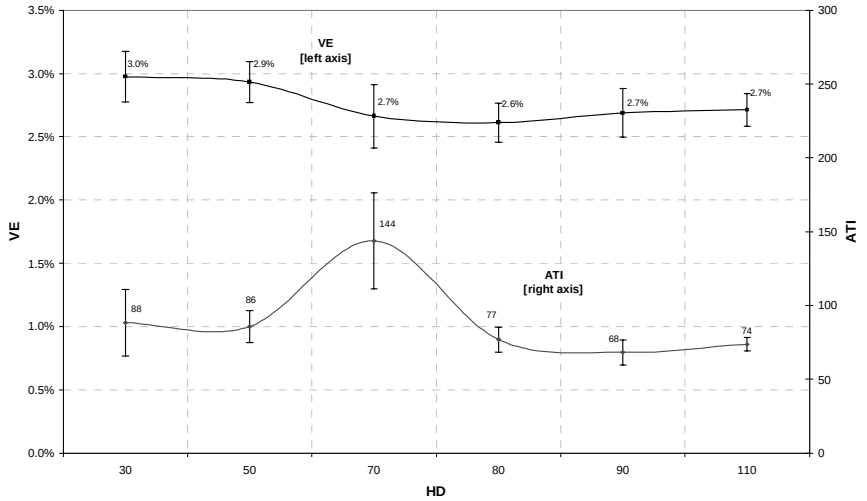

Figure 2: VE (left axis) and ATI (right axis) versus the number of HN (A-LR with $\eta_{ij} \in [0.1, 0.01]$). The vertical bars at each data point indicate the standard deviation.

ESANN'2002 proceedings - European Symposium on Artificial Neural Networks
Bruges (Belgium), 24-26 April 2002, d-side publi., ISBN 2-930307-02-1, pp. 289-294

In Figure 3, we present the comparison among the best results of the F-LR and the A-LR experiments. These results show that when HD = 80, the VE error is minimum both for the F-LR and the A-LR algorithm. The ATI is always lower when considering the A-LR algorithm.

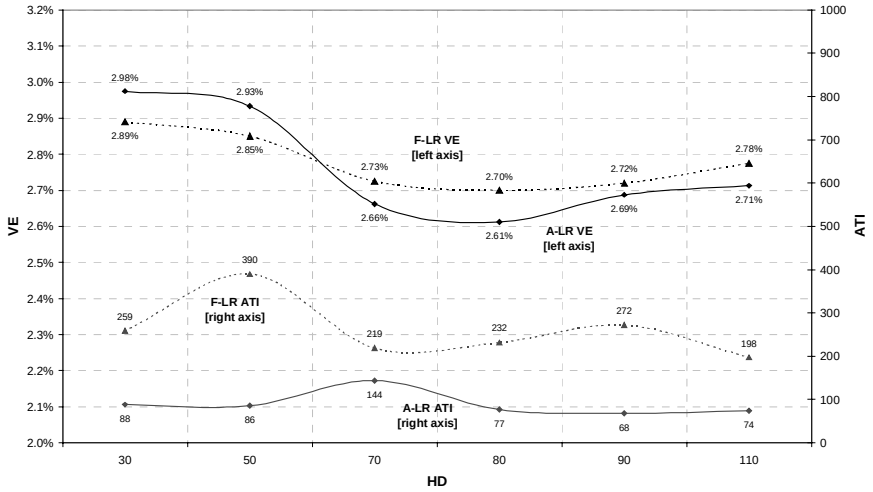

Figure 3: VE (left axis) and ATI (right axis) versus the number of HD. Dotted lines refer to the F-LR algorithms and solid lines refer to the A-LR one. Upper part refers to the VE; lower part refers to the ATI.

We refer to the Rejection Rate (RR) as the percentage of patterns that are not classified (i.e. rejected) according to a given confidence criterion (e.g., a pattern is classified if the difference between the two highest neuron outputs is higher than a confidence threshold).

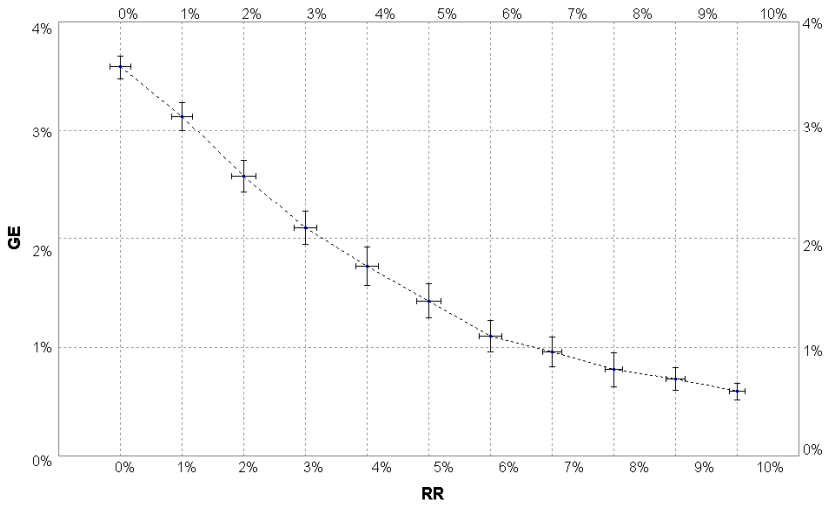

Figure 4: GE versus RR with A-LR algorithm ($\eta_{ij} \in [0.1, 0.01]$).

The GE is computed only on the classified (i.e., not rejected) patterns.
The following results have been obtained by using the test set, A-LR with $\eta_{ij} \in [0.1, 0.01]$ and HD = 80 (see Fig. 2). In Figure 4 the GE versus the RR is shown: the standard deviations both for GE and RR are reported. The results compare favorably with those presented in [5].

4. Conclusions

The aim of our experiments is to validate the BP learning algorithm with a local and adaptive management of the learning rate. We adopt as case study the classification of hand-written characters. We compared the performance with those of the constant learning rate management.
At the beginning, we deeply investigated the network topologies and the learning parameters space. In the best case, the A-LR algorithm exhibits better performance in terms of VE and ATI than the F-LR one. Moreover the GE of the A-LR algorithm gives performance which compare favourably with what reported in [5].
The following table summarizes the best results obtained in the experiments.

	VE		ATI		GE	
	VE	**St. Dev.**	**ATI**	**St. Dev.**	**GE**	**St. Dev.**
F-LR	2.70%	0.14%	232	7.8	3.73%	0.17%
A-LR	2.61%	0.18%	77	1.5	3.58%	0.13%

References

[1] L. Tarassenko, *A Guide to Neural Computing Applications*, Arnold Publishers, London, 1999.

[2] G. M. Bo, H. Chiblè, D. D. Caviglia and M. Valle, *Analog VLSI On-Chip Learning Neural Network with Learning Rate Adaptation*, in Learning on Silicon – Adaptive VLSI Neural Systems, G. Cauwenberghs, M. A. Bayoumi Eds, The Kluwer International Series In Engineering And Computer Science Volume 512, Kluwer Academic Publishers, printed in the Netherlands, June 1999, pp. 305 - 330 (ISBN: 0-7923-8555-1).

[3] P. J. Grother, *NIST Special Database 19*, National Institute of Standards and Technology, 1995

[4] G. M. Bo, D. D. Caviglia, M. Valle, *An Analog VLSI Implementation of a feature Extractor for Real Time Optical Character Recognition*, IEEE Journal of Solid State Circuits, Vol. 33, No 4, pp. 556 – 564, April 1998.

[5] S. Cho, *Neural-Networks Classifiers for Recognizing Totally Uncostrained Handwritten Numerals*, IEEE Transactions on Neural Networks, Vol. 8, No. 1, Jan. 1997