ESANN'1995 proceedings - European Symposium on Artificial Neural Networks
Brussels (Belgium), 19-20-21 April 1995, D-Facto public., ISBN 2-9600049-3-0, pp. 347-352

Performance analysis of a MLP weight initialization algorithm

Mohamed Karouia[1,2], Régis Lengellé[1] and Thierry Denœux[1]

[1] Université de Compiègne – U.R.A. CNRS 817 Heudiasyc
BP 649 - F-60206 Compiègne cedex - France
Tel. (33) 44 23 44 23, Fax. (33) 44 23 44 77
mkarouia@hds.univ-compiegne.fr

[2] Lyonnaise des Eaux (LIAC)

Abstract. The determination of the initial weights is an important issue in multilayer perceptron design. Recently, we have proposed a new approach to weight initialization based on discriminant analysis techniques. In this paper, the performances of multilayer perceptrons (MLPs) initialized by non-parametric discriminant analysis are compared to those of randomly initialized MLPs using several synthetic and real-world benchmark learning tasks. Simulation results confirm that the proposed scheme yields a better initial state, as compared to randomly initialized MLPs. This leads to an improvement in the generalization performance and a reduction in training time, especially for complex or ill-posed problems.

1 Introduction

Backpropagation is one of the most popular training algorithms for MLPs. However, it can be very slow in complex learning tasks. Over the last years, many approaches have been developed in order to speed up the backpropagation algorithm. One such approach consists in finding ways of providing the BP algorithm with as good an initial state as possible. Several methods have been proposed, such as the use of prototypes [2]. Recently, we have proposed a new method for initializing the weights in MLPs, based on discriminant analysis (DA) techniques [6]. Experiments have shown that, at least in some cases, the proposed scheme allows better generalization and a reduction of training time, as compared to random initialization. This study aims at carrying out a more thorough experimental exploration of the proposed initialization procedure, using miscellaneous synthetic and real-world datasets, in order to validate the previous results.

The rest of this paper is organized as follows. Section 2 starts with the description of the weight initialization method. The datasets are described in section 3. Simulation results are presented in section 4 and discussed in section 5.

2 Description of the method

The initialization method concerns feedforward networks with one hidden layer and one output layer of sigmoidal units. The network is trained using the BP algorithm with adaptive learning rates. The weight vectors in the hidden layer

ESANN'1995 proceedings - European Symposium on Artificial Neural Networks
Brussels (Belgium), 19-20-21 April 1995, D-Facto public., ISBN 2-9600049-3-0, pp. 347-352

are determined as discriminant vectors generated by discriminant analysis, of dimension equal to the number d of inputs, and bias terms. The principle of this weight initialization method can be described as follows:

1. **Choose one DA technique.**
2. **Extract the n_h best discriminant vectors $\tau_l, l = 1, \ldots, n_h$**
3. **Initialize the weights of the first hidden layer as $w_{l\bullet} = [\alpha\tau_l, b_l]$ where α is a control parameter and b_l is the bias weight for unit l (at this stage, b_l has an arbitrary value).**
4. **Determine a value of b_l that maximizes a measure of class separability in the space spanned by the hidden units.**
5. **Initialize randomly the hidden-to-output weights.**
6. **Train the output units until no significant error reduction occurs.**
7. **Start the learning process of the whole network.**

In this study, the determination of discriminant vectors in step 2 is done by a particular DA technique called *non-parametric discriminant analysis* (NPDA). It consists in determining the vectors which maximize a nonparametric version of the Fisher criterion defined as $\mathcal{J}(\tau) = \frac{\tau^T B \tau}{\tau^T W \tau}$. τ is a d-dimensional vector on which data are projected. B is nonparametric between-class scatter matrix and W is the parametric within-class scatter matrix [4]. NPDA attempts to preserve the local data structure along the Bayes classification boundary. If the data distributions are significantly non-normal, it allows to find a mapping transformation that preserves the complex structure needed for classification. The number of discriminant vectors extracted by NPDA is not linked to the number of classes and is generally equal to the number of inputs. In step 4, the measure of class separability is defined as $\mathrm{tr}(G^{-1}B)$, where G and B are respectively the total and the between-class covariance matrices of activations in the hidden layer, as suggested in [7]. A more detailed description of the initialization method can be found in [6].

3 Datasets

In order to compare the performances of networks initialized randomly and using our procedure, we have considered a set of 2 real-world and 5 artificial benchmark data sets, the main characteristics of which are summarized in table 1. The data sets T_1 to T_4 were generated from a family of parameterized normal data, as suggested in [3]. These tasks were chosen to investigate the efficiency of the method in the case of ill-posed problems (with few examples as compared to data dimensionality). All the classes have diagonal covariance matrices. The diagonal elements are defined by $D_i(a, b) = a + (b - a)\frac{i-1}{d-1}$, where $a < b$, $i = 1, \ldots, d$. Parameters a and b were fixed to 1 and 10, respectively. For the four tasks, the mean vectors were $m_1 = (0, \ldots, 0)$, $m_2 = (1, \ldots, 1)$ and $m_3 = (1, -1, \ldots, 1, -1)$ for the three classes, respectively.

ESANN'1995 proceedings - European Symposium on Artificial Neural Networks
Brussels (Belgium), 19-20-21 April 1995, D-Facto public., ISBN 2-9600049-3-0, pp. 347-352

Table 1: dataset description

task	input dimension	# classes	# training samples	# test samples	references
T_1	10	3	30	3000	[3]
T_2	10	3	300	3000	"
T_3	30	3	60	3000	"
T_4	30	3	900	3000	"
waveform	21	3	300	300	[1]
vowel	10	11	528	462	[8]
sonar	60	2	104	104	[5]

Table 2: misclassification rates

data sets	initialization technique	number of hidden units	error rates (std) (%) training	test
vowel	NPDA	7	11.1 (0.7)	43 (2.9)
	Random	5	17.9 (2.9)	51.4 (5.6)
sonar	NPDA	12	0 (0)	9.81 (1.1)
	Random	32	1 (0.7)	15.6 (3.6)
waveform	NPDA	4	0.1 (0.1)	15.6 (0.4)
	Random	10	0.4 (0.4)	18.4 (1.3)

4 Results

For each classification task, we varied the number of hidden units n_h from 2 to n_h^{max} ($n_h^{max} \leq d$). The weights were initialized as explained in section 2, and with random numbers. At each learning cycle (epoch), the misclassification rate (the percentage of misclassified examples) was computed over the training and test data sets. For each number of hidden units and for each weight initialization method, the algorithm was run 10 times. The mean misclassification rates as a function of time were computed over the 10 trials, for each value of n_h. Figures 1(a)-(c) show examples of the mean classification rates as a function of time with a given n_h for three of the classification tasks. The minimal value of the mean misclassification rate was also deduced for each n_h and for each weight initialization method. Figures 2(a)-(c) show the evolution of minimal mean misclassification rates as a function of the number of hidden units n_h. Finally, the mean and the standard deviation of the error rates corresponding to the optimal number of hidden units for each method are given in table 2. In the case of normally distributed data, in addition to the mean misclassification rates, we computed a 95% confidence interval on the error probability for each task T_i (i=1,..., 4) and for each n_h value (figures 3(a)-(d)).

ESANN'1995 proceedings - European Symposium on Artificial Neural Networks
Brussels (Belgium), 19-20-21 April 1995, D-Facto public., ISBN 2-9600049-3-0, pp. 347-352

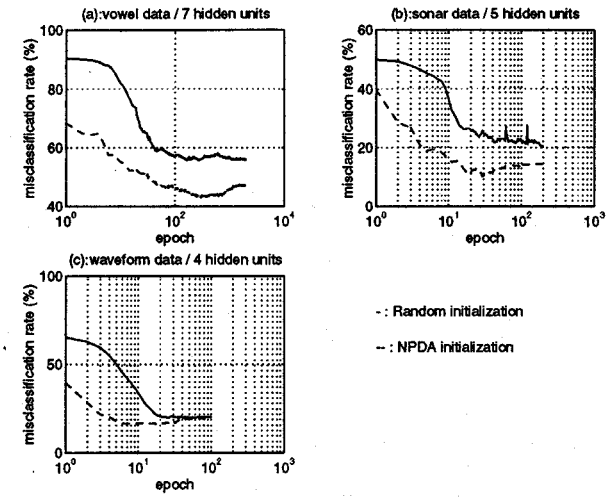

Figure 1: Mean test misclassification rate as a function of training time (averages over 10 trials).

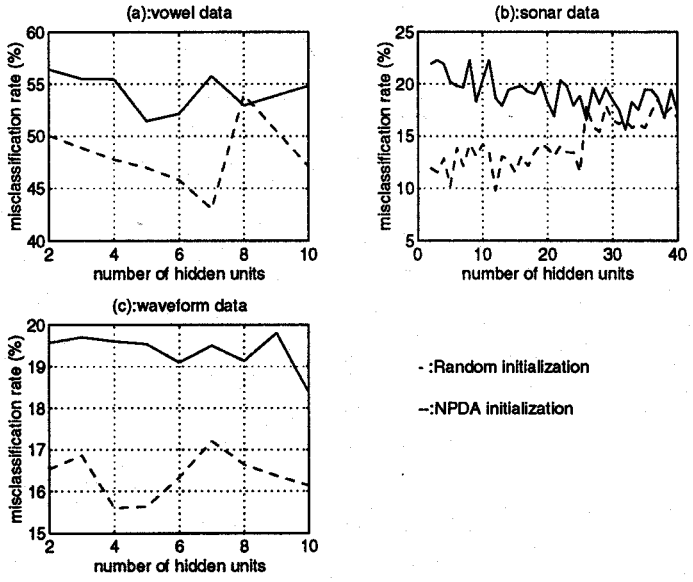

Figure 2: Mean test misclassification rate as a function of the number of hidden units (averages over 10 trials).

ESANN'1995 proceedings - European Symposium on Artificial Neural Networks
Brussels (Belgium), 19-20-21 April 1995, D-Facto public., ISBN 2-9600049-3-0, pp. 347-352

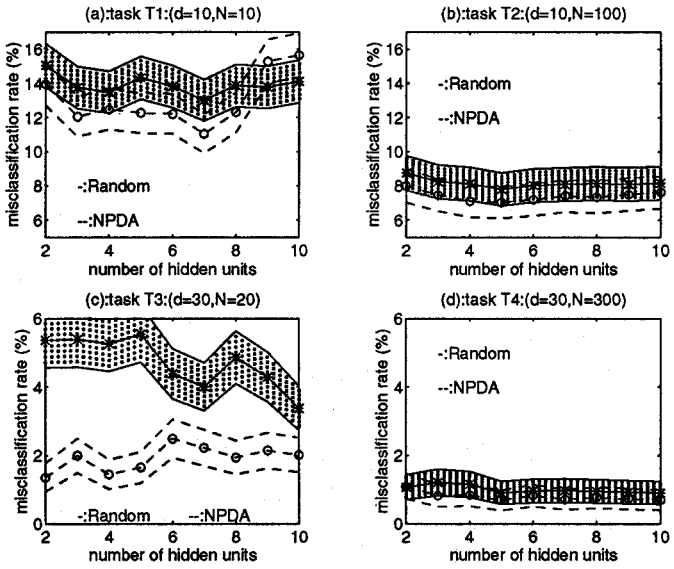

Figure 3: Mean test misclassification rate and a 95% confidence interval as a function of the number of hidden units (averages over 10 trials).

5 Discussion

In the vowel, sonar and waveform datasets, the number of learning samples is relatively large as compared to dimensionality. In these experiments, the initial misclassification rates of the networks initialized with discriminant vectors were always smaller than those of the randomly initialized networks (figures 1(a)-(c)). This better starting state resulted in faster convergence, and was also associated with better generalization (figures 1 and 2). In the examples reported in figures 1(a)-(c), the numbers of training cycles needed to obtain minimal test error rates were respectively 272, 29 and 18 for the vowel, sonar and waveform datasets with NPDA initialization, against 296, 190 and 60 for the same tasks with random initialization. Additionally, the least generalization errors in the sonar and waveform recognition tasks were obtained with significantly fewer hidden units following initialization with NPDA (figure 2).

For the second group of data (T_1 to T_4), the initial misclassification rates were also found to be smaller with NPDA initialization than with random initialization. However, the gain in generalization error was only significant in tasks T_1 and T_3, for which the number of samples is very small as compared to the number of inputs. The results shown in figures 3(a)-(c) show that the gain is also more important in task T_3 than in task T_1. One conclusion that may be drawn from these results is that the gain in generalization ability obtained by initializing MLPs with NPDA seems to be particularly important when the number of samples is small as compared to data dimensionality.

ESANN'1995 proceedings - European Symposium on Artificial Neural Networks
Brussels (Belgium), 19-20-21 April 1995, D-Facto public., ISBN 2-9600049-3-0, pp. 347-352

6 Conclusion

A scheme for initializing the hidden layer weights in MLPs has been studied on two real world and five synthetic classification tasks. This method consists in using discriminant vectors extracted by NPDA as the initial weight vectors in the hidden layer. Simulation results highlighted four advantages of the NPDA initialization method as compared to random initialization: better generalization, especially in complex problems, acceleration of convergence during the learning process, smaller number of hidden units and smaller sensitivity of generalization error to the number of training samples. The proposed scheme allows to find good initial weights for a given number of hidden units. To find the minimal number of hidden units, one has to try several configurations and keep the best one. Coupling this scheme to a strategy for optimal MLP construction is the subject of our current research.

Acknowledgment: The vowel and sonar recognition datasets were retrieved from the UCI repository of machine learning databases.

References

[1] L. Breiman, J. H. Friedman, R. A. Olshen, and C. J. Stone. *Classification and Regression Trees.* Wadsworth, Belmont, CA, 1984.

[2] T. Denœux and R. Lengellé. Initializing back-propagation networks with prototypes. *Neural Networks*, 6(3), 1993.

[3] J. H. Friedman. Regularized discriminant analysis. *J. Am. Statist. Ass.*, 84:165–175, 1989.

[4] K. Fukunaga. *Introduction to statistical pattern recognition.* Electrical Science. 2nd. edition, Academic Press, 1990.

[5] R. P. Gorman and T. J. Sejnowski. Analysis of hidden units in a layered network trained to classify sonar targets. *Neural Networks*, 1:75–89, 1988.

[6] M. Karouia, R. Lengellé, and T. Denœux. Weight initialization in BP networks using discriminant analysis techniques. In *Proceedings of Neuronîmes'94*, 1994.

[7] R. Lengellé and T. Denœux. Optimizing multilayer networks layer per layer without back-propagation. In Igor Aleksander and John Taylor, editors, *Artificial Neural Networks II*, pages 995–998. North-Holland, Amsterdam, 1992.

[8] A. J. Robinson. *Dynamic error propagation networks.* PhD thesis, Cambridge University Engineering, 1989.