IOPscience iopscience.iop.org

Home Search Collections Journals About Contact us My IOPscience

Training of optimal cluster separation networks

This article has been downloaded from IOPscience. Please scroll down to see the full text article.

1994 J. Phys. A: Math. Gen. 27 L387

(http://iopscience.iop.org/0305-4470/27/11/007)

View the table of contents for this issue, or go to the journal homepage for more

Download details:
IP Address: 171.66.16.68
The article was downloaded on 02/06/2010 at 02:15

Please note that terms and conditions apply.

J. Phys. A: Math. Gen. 27 (1994) L387–L390. Printed in the UK

LETTER TO THE EDITOR

Training of optimal cluster separation networks

A Wendemuth

Department of Physics, Theoretical Physics, Oxford University, 1 Keble Road, Oxford OX1 3NP, UK

Received 3 February 1994

Abstract. Finding the optimal separation of two clusters of normalized vectors corresponds to training thresholds and weights in a neural network of maximum stability. In order to achieve this, two local iterative algorithms are presented which treat threshold and weights all in one, avoiding the need to calculate any intermediate 'test' quantities. Convergence is proved, and the separation/stability obtained is shown to match theoretical predictions and to be superior to existing algorithms.

Linear separation of two clusters of vectors has been studied intensively. In particular, in the context of neural networks, supervised learning algorithms have been presented which aim to produce a positive gap size or stability between the two clusters of output. Initially, the Hebb rule leads to the well understood Hopfield model. This model has been shown to yield low or even negative stability of the embedding of patterns (Hertz *et al* 1991). Therefore, algorithms have been proposed which overcome this drawback. The Adaline algorithm will always yield positive stability for loading capacities α less than one (Diederich and Opper 1987, Widrow and Hoff 1960). The minover (Krauth and Mezard 1987) and AdaTron (Anlauf and Biehl 1989) algorithms give the optimal (i.e. best obtainable) stability for networks without thresholds.

However, these algorithms do not achieve maximum gap size/stability since they either fail to treat the threshold at all or just treat the 'threshold dimension' as an augmented one. The threshold can be treated correctly by optimization-theory methods (Fletcher 1987). However, these require inspection of a limited number of 'test solutions' which have to be computed by lengthy matrix inversions. Algorithms have been proposed which do this very economically (Rujan 1993) but which remain algebraically complex non-local procedures. The 'test solutions' for biased patterns may as well be computed by fast iterative schemes (Wendemuth and Sherrington 1993). It is desirable from a computational point of view, as well as from the biological concept of neural networks, not to inspect test solutions at all but to have simple and local iterative algorithms of the kind proposed here. Furthermore, it was shown that optimal stability in networks with adjustable threshold leads to greatly improved generalization ability (Wendemuth 1994b).

The optimal-stability problem can be stated in neural networks and in cluster separation. In neural-network terms, the given problem is to find an N-dimensional perceptron vector J and a threshold T such that for a given set of N-dimensional patterns $\{\xi_1, \ldots, \xi_p\}$ and corresponding outputs ('targets') $\{\tau_1, \ldots, \tau_p\}$ ($\xi_\mu \in R^N$; $|\xi_\mu|^2 = N \quad \forall \mu$; $\tau_\mu \in \{\pm 1\}$), the equations

$$\tau_\mu = \mathrm{sign}(J \cdot \xi_\mu - T) \qquad \mu = 1 \ldots p \tag{1}$$

0305-4470/94/110387+04$19.50 © 1994 IOP Publishing Ltd

are satisfied with maximum stability $\Delta = \Delta_{\text{opt}}$, i.e.

$$\Delta_{\text{opt}} = \min_{\mu}(\Delta_\mu) := \min_{\mu}\left(\frac{J\cdot\tau_\mu\xi_\mu - T\tau_\mu}{|J|}\right) = \max_{\{J',T'\}}\min_{\mu}\left(\frac{J'\cdot\tau_\mu\xi_\mu - T'\tau_\mu}{|J'|}\right). \tag{2}$$

In geometric terms, the two sets of examples belonging to the two classes of output form two *clusters*. The task then is to separate these clusters such that the gap between their convex hulls becomes maximal. The gap size is exactly $2\Delta_{\text{opt}}$, the normal direction from one convex hull to the other is J and the centre of the gap is at distance T from the origin. The hyperplane which is the centre of the gap is then given by all points ξ satisfying $J\cdot\xi - T = 0$.

The main concept of the following two algorithms is to *completely eliminate the threshold from the iteration.* The minimal-overlap algorithm, generalizing the idea of the minover algorithm by Krauth and Mezard (1987), proceeds as follows.

The minimal-overlap algorithm. Define $\sigma_\mu := \tau_\mu\xi_\mu$. Start with $J^{(0)} = \mathbf{0}$ and at any iteration step t find two quantities $\sigma_+^{(t)}$, $\sigma_-^{(t)}$ given by $J^{(t)}\cdot\sigma_\pm^{(t)} = \min_\mu\{J^{(t)}\cdot\sigma_\mu|\tau_\mu = \pm 1\}$.

If $J^{(t)}\cdot\frac{1}{2}(\sigma_+^{(t)} + \sigma_-^{(t)}) \leqslant c$ (c = some fixed positive number) then

$$J^{(t+1)} = J^{(t)} + \frac{1}{N}\frac{1}{2}(\sigma_+^{(t)} + \sigma_-^{(t)})$$

else

$$\text{stop (after } t = M \text{ steps) and set } T = J^{(M)}\cdot\frac{1}{2}(\sigma_+^{(M)} - \sigma_-^{(M)}).$$

The possibility of eliminating the threshold can be understood as follows. Consider $J^{(t)}$ at iteration step t and let the threshold be a free parameter. One would like to use T to find the best possible stability at given (fixed) $J^{(t)}$. After equation (2), one has to find

$$f_{\text{opt}}^{(t)} =: \max_T\min_\mu f_\mu^{(t)}(T) =: \max_T\min_\mu(J^{(t)}\sigma_\mu - T\tau_\mu) \qquad (\mu = 1\ldots p). \tag{3}$$

Keeping the patterns with different outputs separate, one obtains for patterns with $\tau_\mu = (+/-)1$

$$f_\mu^{(t)}(T) \geqslant f_{+/-}^{(t)}(T) =: J^{(t)}\cdot\sigma_{+/-}^{(t)} - (+/-)T \qquad \text{(for any } T) \tag{4}$$

and $f_{+/-}^{(t)}(T)$ is.decreasing/increasing with T. Therefore, $f_{\text{opt}}^{(t)}$ is given by $f_{\text{opt}}^{(t)} = f_-^{(t)}(T) = f_+^{(t)}(T)$ which yields

$$T_{\text{opt}}^{(t)} = J^{(t)}\cdot\tfrac{1}{2}(\sigma_+^{(t)} - \sigma_-^{(t)}) \tag{5}$$

and

$$f_{\text{opt}}^{(t)} = J^{(t)}\cdot\sigma_\pm^{(t)} \mp T_{\text{opt}}^{(t)} = \tfrac{1}{2}J^{(t)}\cdot(\sigma_+^{(t)} + \sigma_-^{(t)}) = \Delta^{(t)}|J^{(t)}|. \tag{6}$$

Equation (5) is used to set the threshold to its optimal value. In equation (6), the optimally set threshold is used and therefore eliminated. Also, from equation (6), the stop condition

of the algorithm is derived. The update turns J into the direction given by equation (6) in the same spirit as in the minover algorithm.

The algorithm is not going to be 'trapped' into a halt at any step, since $\sigma_+^{(t)} + \sigma_-^{(t)} = \mathbf{0}$ will never be obtained. If it were, one would have had $\xi_+^{(t)} = \xi_-^{(t)}$. This would mean that the same pattern would be mapped to $+1$ and to -1 simultaneously, which by feasibility of the solution cannot occur.

The proof of this algorithm is achieved by mapping it onto the proof of the minover algorithm. The first alteration is that updates are in the space of 'difference patterns' v_μ, defined by $v_\mu = 0.5[\sigma_\mu^+ + \sigma_\mu^-]$. Note that if there are (kp) patterns with positive output, the total number of difference patterns is $D = p^2k(1-k)$. However, this does not cause an increase in convergence time: if one compares the standard minover algorithm without threshold to the minimal-overlap algorithm, the latter detects, at every sweep t, the quantities $\sigma_+^{(t)}$ and $\sigma_-^{(t)}$ *simultaneously* by inspecting the p patterns. The number of update sweeps in order to achieve a given fraction F of the optimal stability remains proportional to $1/(1-F)$ as known from Wendemuth *et al* (1993).

The second alteration is the treatment of the threshold as a dependent variable. With T optimally set at any iteration step, the least stable patterns are $\sigma_+^{(t)}$ and $\sigma_-^{(t)}$. Adding the positive and negative terms in equations (4) and using equation (3), one eliminates the threshold and obtains $J^{(t)} \cdot v_\mu \geqslant 0.5 J^{(t)} \cdot (\sigma_+^{(t)} + \sigma_-^{(t)})$ $(\mu = 1 \ldots D)$. Therefore, one may regard patterns in difference spaces throughout. In particular, $J^* \cdot v_\mu \geqslant J^* \cdot 0.5(\sigma_+^* + \sigma_-^*) \geqslant c = \Delta_{\text{opt}} |J^*|$ $(\mu = 1 \ldots D)$. J^* is the weight vector of the optimally-stable network. It is assumed that such a network exists with $\Delta_{\text{opt}} > 0$.

If one uses the proof from Krauth and Mezard (1987), replacing their σ_μ by v_μ, convergence follows on noting that $\sqrt{N} \geqslant |v_\mu| > 0$ $(\mu = 1 \ldots D)$. For $c \to \infty$, the algorithm will converge to optimal stability. Clearly, this algorithm works whatever the correlation between the patterns is.

In the restricted gradient-descent algorithm, updates are again performed with *all* difference patterns v_μ. The perceptron vector can be written in terms of the *embedding strengths* x_μ^+ and x_μ^- as $J = \sum_{\mu=1}^{kp} x_\mu^+ \sigma_\mu^+ + \sum_{\mu=1}^{(1-k)p} x_\mu^- \sigma_\mu^-$. One then proceeds as follows.

The restricted gradient-descent algorithm. Start with $J(0) = \mathbf{0}$ and update J sequentially with all difference patterns according to

$$\delta x_\mu^+ = \delta x_\mu^- = \max \left\{ \left[\frac{\gamma}{N} (1 - J(t) \cdot v_\mu) \right], -x_\mu^+, -x_\mu^- \right\} \tag{7}$$

where $0 < \gamma < 1$. Stop if for the last set $(\mu = 1 \ldots D)$ of updates

$$\max_{\mu=1 \ldots D} |\delta x_\mu^\pm| < \varepsilon \tag{8}$$

and set T as in equation (5).

The stop condition will be met after a finite number of updates for any ε and, for $\varepsilon \to 0$, optimal stability is reached.

Since gradient-descent steps are performed, the difference between the obtained stability and the optimal one will decay exponentially. Decay times can be computed as in Wendemuth *et al* (1993), showing that the number of learning sweeps required for random input patterns is $\propto \ln(1/(1-F))$ which is much smaller than in the minimal-overlap algorithm. However, one has to consider that one learning sweep contains $D \propto p^2$ update

steps whereas in the minimal-overlap algorithm, the two quantities $\sigma_+^{(t)}$ and $\sigma_-^{(t)}$ in any sweep are found by stepping through the p patterns once. The restricted gradient-descent algorithm will therefore be faster than the minimal-overlap algorithm only for small p and if high accuracies F are desired. Note that this difference in sweep sizes did not occur in the respective algorithms without thresholds since, then, every sweep contained p steps.

The proof of the algorithm can again be mapped to the AdaTron proof (Anlauf and Biehl 1989) applied to the space of difference patterns. Since by construction $\delta x_\mu^+ = \delta x_\mu^-$, it is ensured that updates are indeed performed with difference patterns only. The algorithm then performs gradient-descent steps in the space of difference patterns, approaching a solution $J = \sum_{\mu=1}^{p} \lambda_\mu v_\mu = \sum_{\mu=1}^{kp} x_\mu^+ \sigma_\mu^+ + \sum_{\mu=1}^{(1-k)p} x_\mu^- \sigma_\mu^-$ with $Jv_\mu \geqslant 1$ $(\mu = 1 \cdots D)$ where the restriction to positive embedding strengths ensures that $\lambda_\mu \geqslant 0$. In analogy with the AdaTron proof, $\lambda_\mu \geqslant 0$ guarantees optimality of the selected subset of patterns which form the perceptron vector. The threshold is restored as in the minover case and is therefore optimal, which completes the proof. The performances of the two algorithms as well as a full proof are given in Wendemuth (1994a) and will be published elsewhere.

It is a pleasure to thank D Sherrington for inspiring discussions. Also, I would like to acknowledge support by the Science and Engineering Research Council of Great Britain, the Friedrich–Naumann–Stiftung and the European Community under contract no ERB4001GT922302.

References

Anlauf J K and Biehl M 1989 *Europhys. Lett.* **10** 687
Diederich S and Opper M 1987 *Phys. Rev. Lett.* **58** 949
Fletcher R 1987 *Practical Methods of Optimization* (New York: Wiley)
Hertz J, Krogh A and Palmer R G 1991 *Introduction to the Theory of Neural Computation* (Redwood City, CA: Addison-Wesley)
Krauth W and Mezard M 1987 *J. Phys. A: Math. Gen.* **20** L745
Rujan P 1993 *J. Physique* **3** 277–90
Wendemuth A 1994a *DPhil Thesis* Oxford University
Wendemuth A 1994b *J. Phys. A: Math. Gen.* **27** 2325–33
Wendemuth A, Opper M and Kinzel W 1993 *J. Phys. A: Math. Gen.* **26** 3165
Wendemuth A and Sherrington D 1993 *Int. J. Neural Systems* **4**
Widrow B and Hoff M E 1960 *Adaptive Switching Circuits (IRE WESCON Convention Report)* vol 4 pp 4–96