Multistate perceptrons: learning rule and perceptron of maximal stability

1992 J. Phys. A: Math. Gen. 25 5039

(http://iopscience.iop.org/0305-4470/25/19/016)

J. Phys. A: Math. Gen. 25 (1992) 5039–5045. Printed in the UK

Multistate perceptrons: learning rule and perceptron of maximal stability

E Elizalde and S Gómez

Departament d'Estructura i Constituents de la Matèria, Facultat de Física, Universitat de Barcelona, Diagonal 647, 08028 Barcelona, Spain

Received 28 February 1992, in final form 19 May 1992

Abstract. A new perceptron learning rule which works with multilayer neural networks made of multistate units is obtained, and the corresponding convergence theorem is proved. The definition of a perceptron of maximal stability is enlarged in order to include these new multistate perceptrons, and a proof of the existence and uniqueness of such optimal solutions is outlined.

1. Introduction

Perceptrons constitute the simplest architecture for a layered feed-forward neural network (Rosenblatt 1958). An input layer feeds the only unit of the second layer, where the output is read. Thus, there are as many weights ω_k as input units—say N—and just one threshold U. The activation function which decides the final state of the output unit is usually taken to be

$$g(h) \equiv \begin{cases} 0 & \text{if } h < 0 \\ 1 & \text{if } h \geqslant 0 \end{cases}$$

where the field h is calculated, as a function of the input pattern $\boldsymbol{\xi}$, through the formula

$$h \equiv \boldsymbol{\omega} \cdot \boldsymbol{\xi} - U.$$

Learning amounts to finding the weights $\boldsymbol{\omega}$ and the threshold U which map a set of input patterns $\{\boldsymbol{\xi}^\mu\}_{\mu=1,\dots,p}$ into their corresponding outputs $\{\zeta^\mu\}_{\mu=1,\dots,p}$. Among all the possible input–output associations, only the so-called *linearly separable* problems have perceptron solutions. The well-known *perceptron convergence theorem* gives a *perceptron learning rule* to obtain these solutions whenever they exist (Minsky and Papert 1969). Moreover, recent works have developed fast converging algorithms to find the perceptron *solution of maximal stability* (e.g. Krauth and Mézard 1987, Ruján 1991).

The binary perceptron divides the input space in two half-spaces, one for each possible value of the output. The problem of classifying in more than two classes with the aid of a collection of perceptrons is well known in the literature (see e.g. Duda and Hart (1973)). Likewise, if the mapping to be learned has a continuous

0305-4470/92/195039+07$04.50

output, it can be related to the previous classification scheme in two steps: partition of the interval of variation of the continuous parameter in a finite number of pieces—to arbitrary precision—and assignment of each one to a certain base 2 vector (see Gallant (1990)). For instance, a 'thermometer' representation for the interval [0,1] could be

$$\zeta = \begin{cases} (0,0,0,0) & \text{for } y \in [0,0.2) \\ (1,0,0,0) & \text{for } y \in [0.2,0.4) \\ (1,1,0,0) & \text{for } y \in [0.4,0.6) \\ (1,1,1,0) & \text{for } y \in [0.6,0.8) \\ (1,1,1,1) & \text{for } y \in [0.8,1] \end{cases} \tag{1.1}$$

which reduces the learning problem to a five class classification. However, even if this four-perceptron network has learned the thermometer-like $\xi^\mu \longmapsto \zeta^\mu$, $\mu = 1,\ldots,p$ correspondence, new inputs supplied to the net may produce ouputs such as $(0,0,1,1)$ or $(1,0,1,0)$, which cannot be interpreted within this representation; in fact, most of the available codifying schemes suffer from the same inconsistency.

One natural way of avoiding these problematic and rather artificial conversions from continuous to binary data is the use of *multistate unit* perceptrons (see e.g. Rieger (1990) and Nadal and Rau (1991)). With them, only the first of the two steps mentioned above is necessary, i.e. the discretization of the continuous interval. Geometrically, multistate units define a vector in the input space which points to the direction of increase of the output parameter, the boundaries being parallel hyperplanes. That is why this method removes meaningless patterns, since this partition clearly incorporates the underlying relation of order.

At first sight it may seem that the structure derived from a set of binary perceptrons is richer than that arising from a single multistate unit. Nevertheless, it must be taken into account that combinations of multistate perceptrons will be needed whenever the learning problem is not *multistate separable*, giving rise to multilayer neural networks made of multistate units.

In this article we shall first introduce a new *multistate perceptron learning rule*, and then prove the corresponding convergence theorem. Then, the concept of a solution with *maximal stability* will be extended to multistate perceptrons, and their most remarkable properties will be stated. Finally, possible applications of the model of multistate neural networks and some open problems will be discussed.

2. Multistate perceptron convergence theorem

A Q-state neuron may be in any one of Q different output values or 'colours' $\sigma_1 < \cdots < \sigma_Q$. They constitute the result of processing an incoming stimulus through an activation function of the form

$$g_U(h) \equiv \begin{cases} \sigma_1 & \text{if } h < U_1 \\ \sigma_v & \text{if } U_{v-1} \leqslant h < U_v \qquad v = 2,\ldots,Q-1 \\ \sigma_Q & \text{if } U_{Q-1} \leqslant h. \end{cases} \tag{2.1}$$

Therefore, $Q-1$ thresholds $U_1 < \cdots < U_{Q-1}$ have to be defined for each updating unit, which, in the case of the perceptron, is reduced to just the output unit. The field now simply reads

$$h \equiv \omega \cdot \xi. \tag{2.2}$$

Let us distribute the input patterns in the following subsets:

$$\mathcal{F}_v \equiv \{\xi^\mu | \zeta^\mu = \sigma_v\} \qquad v = 1, \ldots, Q. \tag{2.3}$$

From a geometrical point of view (Ruján 1990) the output processor corresponds to the set of $Q - 1$ *parallel hyperplanes*

$$\omega \cdot \xi = U_v \qquad v = 1, \ldots, Q - 1 \tag{2.4}$$

which divide the input space into Q ordered regions, one for each of the colours $\sigma_1, \ldots, \sigma_Q$. Thus, the map $\xi^\mu \longmapsto \zeta^\mu$, $\mu = 1, \ldots, p$, is said to be *learnable* or *separable* if it is possible to choose hyperplanes such that each $\mathcal{F}_v$ be in the zone of colour σ_v.

This picture makes us realize that the fundamental parameters to be searched for while learning are the components of the unit vector $\hat{\omega}$ and *not* the thresholds, since these can be assigned a value as follows. If the input–output map is learnable then

$$\zeta^\mu = g_U(\omega \cdot \xi^\mu) \qquad \mu = 1, \ldots, p \tag{2.5}$$

yields

$$\left.\begin{array}{l} \forall \xi_v^\rho \in \mathcal{F}_v \\ \forall \xi_{v+1}^\gamma \in \mathcal{F}_{v+1} \end{array}\right\} \Longrightarrow \omega \cdot \xi_v^\rho < \omega \cdot \xi_{v+1}^\gamma \tag{2.6}$$

which means that, defining

$$\begin{array}{l} \xi_v^\alpha \in \mathcal{F}_v | \omega \cdot \xi_v^\alpha \geqslant \omega \cdot \xi_v^\rho \ \forall \xi_v^\rho \in \mathcal{F}_v \\ \xi_v^\beta \in \mathcal{F}_v | \omega \cdot \xi_v^\beta \leqslant \omega \cdot \xi_v^\gamma \ \forall \xi_v^\gamma \in \mathcal{F}_v \end{array} \tag{2.7}$$

we get

$$U_v \in \left] \omega \cdot \xi_v^\alpha \omega \cdot \xi_{v+1}^\beta \right] \qquad v = 1, \ldots, Q - 1. \tag{2.8}$$

Hence, during the learning process it is possible to choose

$$U_v = \frac{\omega \cdot \xi_v^\alpha + \omega \cdot \xi_{v+1}^\beta}{2} \qquad v = 1, \ldots, Q - 1 \tag{2.9}$$

which is the best choice for the thresholds with the given ω. Here lies the difference between our approach and that of recent papers such as (Mertens *et al* 1991), where the thresholds are compelled to be inside certain intervals given beforehand. Consequently, we have somehow enlarged their notion of learnability.

Our proposal for the *multistate perceptron learning rule* stems from the following theorem.

Theorem. If there exists ω^* such that $\omega^* \cdot \xi_v^\rho < \omega^* \cdot \xi_{v+1}^\gamma$ for all $\xi_v^\rho \in \mathcal{F}_v$ and $\xi_{v+1}^\gamma \in \mathcal{F}_{v+1}$, $v = 1, \ldots, Q-1$, then the program

```
Start choose any value for ω and η > 0;
Test  choose v ∈ {1,...,Q−1}, ξ_v^ρ ∈ F_v and ξ_{v+1}^γ ∈ F_{v+1};
      if ω·ξ_v^ρ < ω·ξ_{v+1}^γ then go to Test
                                else go to Add;
Add   replace ω by ω + η(ξ_{v+1}^γ − ξ_v^ρ);
      go to Test.
```

will go to Add only a finite number of times.

Corollary. The previous algorithm finds a multistate perceptron solution to the map $\xi^\mu \longmapsto \zeta^\mu$, $\mu = 1, \ldots, p$ whenever it exists, provided the maximum number of passes through Add is reached. This may be achieved by continuously choosing pairs $\{\xi_v^\rho,\ \xi_{v+1}^\gamma\}$ such that $\omega \cdot \xi_v^\rho \geqslant \omega \cdot \xi_{v+1}^\gamma$.

Proof. Define

$$G(\omega) \equiv \frac{\omega \cdot \omega^*}{\|\omega\|\|\omega^*\|} \leqslant 1 \tag{2.10}$$

$$\delta \equiv \min_{v,\rho,\gamma} \left(\omega^* \cdot \xi_{v+1}^\gamma - \omega^* \cdot \xi_v^\rho\right) > 0 \tag{2.11}$$

$$M^2 \equiv \max_{v,\rho,\gamma} \left\|\xi_{v+1}^\gamma - \xi_v^\rho\right\|^2 > 0. \tag{2.12}$$

On successive passes of the program through Add,

$$\omega^* \cdot \omega_{t+1} \geqslant \omega^* \cdot \omega_t + \eta\delta \tag{2.13}$$

$$\|\omega_{t+1}\|^2 \leqslant \|\omega_t\|^2 + \eta^2 M^2. \tag{2.14}$$

Therefore, after n applications of Add,

$$G(\omega_n) \geqslant L(\omega_n) \tag{2.15}$$

$$L(\omega_n) \equiv \frac{\omega^* \cdot \omega_0 + n\eta\delta}{\|\omega^*\|\sqrt{\|\omega_0\|^2 + n\eta^2 M^2}} \tag{2.16}$$

which, for large n, goes as

$$L(\omega_n) \approx \sqrt{n}\frac{\delta}{\|\omega^*\| M}. \tag{2.17}$$

However, n cannot grow at will since $G(\omega) \leqslant 1$, $\forall\omega$, which implies that the number of passes through Add has to be finite.

It is interesting to note that no assumption has been made on the number and nature of the input patterns. Thus, the theorem applies even when an infinite number of pairs of patterns is present and also to inputs not belonging to the 'lattice' $\{\sigma_1, \ldots, \sigma_Q\}^N$.

3. Multistate perceptron of maximal stability

In the previous section an algorithm for finding a set of parallel hyperplanes which separates the $\mathcal{F}_v$ sets in the correct order has been found, under the assumption that such solutions exist. The problem we are going to address now is that of selecting the 'best' of all such solutions.

It is our precise prescription that the *multistate perceptron of maximal stability* has to be defined as the one whose smallest gap between the pairs $\{\mathcal{F}_v, \mathcal{F}_{v+1}\}$, $v = 1, \ldots, Q-1$ is maximal. These gaps are given by the numbers

$$R_v(\omega) \equiv \min_{\rho,\gamma} \left(\frac{\omega}{\|\omega\|} \cdot (\xi^\gamma_{v+1} - \xi^\rho_v) \right) = \frac{\omega}{\|\omega\|} \cdot \left(\xi^\beta_{v+1} - \xi^\alpha_v \right) \tag{3.1}$$

where to obtain the second expression we have made use of the definitions in (2.7). Therefore, calling $\mathcal{D} \subset \boldsymbol{R}^N$ the set of all the solutions to the multistate perceptron problem, the function to be maximized is

$$R(\omega) \equiv \begin{cases} \min\limits_{v=1,\ldots,Q-1} R_v(\omega) & \text{if } \omega \in \mathcal{D} \\ 0 & \text{if } \omega \notin \mathcal{D}. \end{cases} \tag{3.2}$$

Since $R(\lambda\omega) = R(\omega)$, $\forall \lambda > 0$, it is actually preferable to restrict the domain of R to the hyper-sphere $S^N \subset \boldsymbol{R}^N$, i.e.

$$\widetilde{R} : S^N \longrightarrow \boldsymbol{R}^N \qquad \text{such that } \hat{\omega} \longmapsto \widetilde{R}(\hat{\omega}) \equiv R(\hat{\omega}). \tag{3.3}$$

The basic properties of $\widetilde{R}$ are as follows.

(i) $\widetilde{R}(\hat{\omega}) > 0 \iff \hat{\omega} \in \mathcal{D} \cap S^N$.

(ii) The set $\mathcal{D}$ is convex.

(iii) The restriction of $\widetilde{R}$ to $\mathcal{D} \cap S^N$ is a strictly concave function.

(iv) The restriction of $\widetilde{R}$ to $\mathcal{D} \cap S^N$ has a unique maximum.

This last property assures the existence and uniqueness of a perceptron of maximal stability, and it is a direct consequence of the preceding propositions. Moreover, it asserts that *no other* relative maxima is present, which is of great practical interest whenever this optimal perceptron has to be explicitly found.

In Mertens *et al* (1991) the optimization procedure constitutes a forward generalization of the AdaTron algorithm (Anlauf and Biehl 1989). Here the situation is much more complicated because the function we want to maximize is not simply quadratic with linear constraints, but a piecewise combination of them due to the previous discrete minimization taken over the gaps. Thus, we have not been able to find a suitable optimization method which could take advantage of the particularities of this problem. The designing of such converging algorithms is an open question which deserves further investigation.

4. Conclusions

A new *perceptron learning rule* which can be used with perceptrons made of *multistate units* has been derived. Its convergence has been proven to be guaranteed whenever the set of input–output patterns is *multistate separable*. This last concept constitutes an extension of the so-called linear separability, which involves a single hyperplane lying in the input space. Moreover, we have generalized the definition of *perceptron of maximal stability* to encompass the case of multistate simple perceptrons, and their main characteristics and properties have been shown. In particular, the existence and uniqueness of such solutions implies that any standard optimizing method can, in principle, be used, but the determination of the best procedure is an open question left to future research.

The comparison between the actual performances of binary and multistate units in multilayer neural networks is of great interest. Nevertheless, it cannot be established in practice until a good multilayer learning method is found, i.e. a learning rule which may be used even when non multistate-separable problems are treated. To be specific, some learning algorithms for binary units do exist (e.g. the 'tiling algorithm' (Mézard and Nadal 1989) and 'sequential learning' (Marchand *et al* 1990)) which add hidden layers and units in such a way that a correct mapping between the binary input and output patterns is always ensured; and their main tool is the repeated use of the 'binary perceptron learning rule'.

Therefore, our first objective will now consist in designing an always-converging multilayer learning rule with hidden multistate units. Its existence is ensured by the fact that it is actually possible to learn the separation of each of the colours from the rest (with the help of two-colour units) and then the resulting hidden representation turns out to be multistate separable; but this is certainly not a good solution because most of the hidden units turn out to be binary.

Acknowledgments

This work has been partly supported by Dirección General de Investigación Científica y Técnica (DGICYT) project no PB90-0022. SG thanks the Spanish Ministry of Education and Science for an FPI fellowship. We are grateful to the referee for very interesting suggestions.

References

Anlauf J K and Biehl M 1989 The AdaTron: an adaptive perceptron algorithm *Europhys. Lett.* **10** 687

Duda R O and Hart P E 1973 *Pattern Classification and Scene Analysis* (New York: Wiley)

Gallant S I 1990 Perceptron-based learning algorithms *IEEE Trans. Neural Networks* **1** 179

Krauth W and Mézard M 1987 Learning algorithms with optimal stability in neural networks *J. Phys. A: Math. Gen.* **20** L745

Marchand M, Golea M and Ruján P 1990 A convergence theorem for sequential learning in two-layer perceptrons *Europhys. Lett.* **11** 487

Mertens S and Köhler H M 1991 and Bos S Learning grey-toned patterns in neural networks *J. Phys. A: Math. Gen.* **24** 4941

Mézard M and Nadal J P 1989 Learning in feedforward layered networks: the tiling algorithm *J. Phys. A: Math. Gen.* **22** 2191

Minsky M and Papert S 1969 *Perceptrons* (Cambridge, MA: MIT Press)

Nadal J P and Rau A 1991 Storage capacity of a Potts-perceptron *J. Physique* **1** 1109

Rieger H 1990 Properties of neural networks with multistate neurons *Statistical Mechanics of Neural Networks (Lecture Notes in Physics 368)* ed L Garrido (Berlin: Springer)

Rosenblatt F 1958 *Psych. Rev.* **62** 386; *Principles of Neurodynamics* (New York: Spartan)

Ruján P A 1990 Learning in multilayer networks: a geometric computational approach *Statistical Mechanics of Neural Networks (Lecture Notes in Physics 368)* ed L Garrido (Berlin: Springer)
—— 1991 A fast method for calculating the perceptron with maximal stability *Preprint* Oldenburg University