Proc. of ICANNGA (Int. Conf. on Artificial Neural Networks and Genetic Algorithms), pp. 126-129, Prague/C.R., 22-25 April 2001.

An Immunological Approach to Initialize Feedforward Neural Network Weights

Leandro Nunes de Castro[1] & Fernando José Von Zuben[1]

Abstract

The initial weight vector to be used in supervised learning for multilayer feedforward neural networks has a strong influence in the learning speed and in the quality of the solution obtained after convergence. An inadequate initial choice may cause the training process to get stuck in a poor local minimum, or to face abnormal numerical problems. In this paper, we propose a biologically inspired method based on artificial immune systems. This new strategy is applied to several benchmark and real-world problems, and its performance is compared to that produced by other approaches already suggested in the literature.

1. Introduction

The importance of a proper choice for the initial set of weights (weight vector) is stressed by Kolen and Pollak [12]. They showed that it is not feasible to perform a global search to obtain the optimal set of weights. So, for practical purposes, the learning rule should be based on optimization techniques that employ local search to find the solution [19]. As an important outcome of their procedure, there is the fact that a local search process results in a solution strongly related to the initial configuration of the weight vector. It happens because each initial condition belongs to the basis of attraction of a particular local optimum in the weight space, to which the solution will converge [8]. Consequently, only a local optimum can be produced as the result of a well-succeeded training process. If such a solution happens to be the global or a good local optimum, the result is a properly trained neural network. Otherwise, an inferior result will be achieved, so that the poorer the local optimum, the worse the performance of the trained neural network.

This correlation between the initial set of weights and the quality of the solution resembles the existing correlation between the initial antibody repertoire and the quality of the response of natural immune systems, that can be seen as a complex pattern recognition device with the main goal of protecting our body from malefic external invaders, called antigens. Antibodies are the primary immune elements that bind to antigens for their posterior destruction by other cells [9]. The number of antibodies contained in our immune system is known to be much inferior to the number of possible antigens, making the diversity and individual binding capability the most important properties to be exhibited by the antibody repertoire. In this paper, we present a simulated annealing approach, called SAND (Simulated ANnealing for Diversity), that aims at generating a dedicated set of weights that best covers the weight space, to be searched in order to minimize the error surface. The strategy assumes no a priori knowledge about the problem, except for the assumption that the error surface has multiple local optima. In this case, a good sampling exploration of the error surface is necessary to improve the chance of finding a promising region to search for the solution. The algorithm induces diversity in a population by maximizing an energy function that takes into account the inverse of the affinity among the antibodies. The weights of the neural network will be associated with antibodies in a way to be further elucidated.

2. The Simulated Annealing Algorithm

The simulated annealing algorithm makes a connection between statistical mechanics and combinatorial optimization [7,10]. The origin of the method is associated with aggregate properties of a large number of atoms found in samples of liquids or solid matters. The behavior of the system in thermal equilibrium, at a given temperature, can be characterized experimentally by small fluctuations around the average behavior. Each atomic position is weighted by a probability factor

$$P(\Delta E) = \exp(-\Delta E / T), \quad (1)$$

where E is the energy of the configuration, T the temperature and ΔE a small deviation in the energy measured. At each step of this algorithm, an atom is given a small random displacement and the resulting change, ΔE, in the energy of the system is computed. If $\Delta E \leq 0$, the displacement is accepted, and the configuration with the displaced atom is used as the starting point of the next step. The case $\Delta E > 0$ is treated probabilistically: the probability of accepting the new configuration is given by Equation (1).

The temperature is simply a control parameter in the same unit as the cost (energy) function. The simulated annealing process consists of first "melting" the system being optimized at a high effective temperature, then lowering the temperature by slow stages until the system "freezes" and no further change occurs (steps of increasing temperature can also be incorporated). At each temperature, the simulation must proceed long enough for the system to reach a steady state. Notice that, transitions out of a local optimum are always possible at nonzero temperatures. The sequence of temperatures and the size of the ΔE variation are considered an annealing schedule.

[1] School of Electrical and Computer Engineering, State University of Campinas, Brazil, e-mail: {lnunes,vonzuben}@dca.fee.unicamp.br

3. An Immunological Approach

The immune system model used in this work is a simplification of the biological one. Real-valued vectors represent the antibodies (Ab). In our problem, the antigen (Ag) population (training set) will be disregarded, so the energy measure of the population of antibodies (set of weights) will be determined based solely on the individuals of the population of antibodies. The binding Ag–Ab represents a measure of complementarity between an antigen and an antibody, an idea that can be adapted to determine the complementarity among members of the antibody repertoire. The goal is to maximize the distance among antibodies (Ab–Ab), with the purpose of reducing the amount of similarities within the population.

An abstract model to describe Ag–Ab interactions was introduced by Perelson & Oster [18]. In this model, it is assumed that the features of an antibody receptor (combining region) relevant to antigen binding can be described by specifying a total of L shape parameters. It is also assumed that the same L parameters can be used to describe an antigen. Combining these L parameters into a vector, the antibody receptors and the antigen determinants can be described as points Ab and $\overline{\text{Ag}}$, respectively, in an L-dimensional Euclidean vector space, called shape-space S. Here we use $\overline{\text{Ag}}$ (the complement of Ag) because the affinity is directly proportional to complementarity, and affinity will be associated with the proximity of Ab to $\overline{\text{Ag}}$. By defining a metric on S, the proximity between Ab and $\overline{\text{Ag}}$ is a measure of their affinity. The antibodies and the complement of the antigens were represented by a set of real-valued coordinates. Thus, mathematically, each molecule could be regarded as a point in an L-dimensional real-valued space, and the affinity Ag–Ab was related to the inverse of the Euclidean distance between them.

Any given antibody is assumed to recognize some set of antigens and therefore covers some portion of the space.

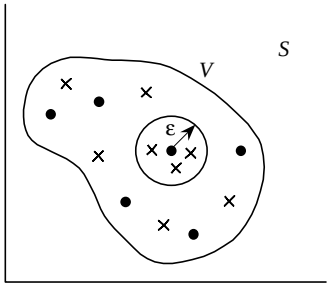

Figure 1: Within shape-space S, there is a volume V in which the antibodies (•) and the complement of antigens (x) are located. An antibody is assumed to be able to bind any complement of antigen within a distance ε (region of stimulation).

In the natural immune system, the binding Ag–Ab might not be fully complementary for the lymphocytes to become activated, since a partial matching might suffice. It implies that, if no error were allowed, the immune cells would become activated only when a perfect match occurred. Based on this argument, it is defined an acceptable matching distance (ε), that determines the coverage provided by the antibodies. Some authors called ε a ball of stimulation [20], because it represents the group of antigens that can stimulate the antibody contained in its center. Figure 1 depicts the shape-space model and the region of stimulation, where the dots and crosses denote the location of antibodies and the complement of antigens, respectively. The circle of radius ε around one of the antibodies shows its coverage. This is an illustrative abstraction, once Ag-Ab recognition happens through shape complementarity, instead of similarity as in $\overline{\text{Ag}}$-Ab recognition.

As we are dealing with real-valued vectors, the inverse of the Euclidean distance can be used as the measure of affinity between the molecules:

$$Aff(x_i, x_j) = \frac{1}{ED(x_i, x_j) + \varepsilon} \tag{2}$$

where x_i and x_j represent independent vectors of length L, ε is a small positive constant, and

$$ED(x_i, x_j) = \sqrt{\sum_{k=1}^{l} (x_{ik} - x_{jk})^2} \ , \tag{3}$$

Assuming an Euclidean search-space, the energy measure to be optimized can be simply defined as the sum of the Euclidean distances among all vectors that represent the antibody population

$$E = \sum_{i=1}^{N} \sum_{j=i+1}^{N} ED(x_i, x_j) \, . \tag{4}$$

A stopping criterion for the simulated annealing algorithm, which takes into account the diversity among the vectors, has to be defined. The approach to be proposed here, among other possibilities, involves the analysis of directional data.

Given the vectors x_i, $i = 1,..., N$, it is initially necessary to transform them into unit vectors, resulting in a set $\{ \bar{x}_i , i = 1,..., N\}$ of unit vectors of length L. The average directional vector is

$$\bar{x} = \frac{1}{N} \sum_{i=1}^{N} \bar{x}_i \ . \tag{5}$$

A metric to estimate the diversity, or equivalently the uniformity, of the distribution of the unit vectors in a hypersphere can be simply given by

$$\left\| \bar{x} \right\| = \left(\bar{x}^T \bar{x} \right)^{1/2} , \tag{6}$$

where $\| \cdot \|$ represents the Euclidean norm.

The stopping criterion (*SC*) is then based on the index

$$I_{SC} = 100 \times (1 - \|\bar{x}\|) . \quad (7)$$

Equation (7) is the percentile norm of the resultant vector $\bar{x}$, and is equal to 100% when this norm is zero. In practical terms, a value of I_{SC} close to 100% for the stopping criterion (*SC*) is a reasonable choice, indicating a distribution close to uniform.

3.1. Neural Network Weights and Antibodies

Each antibody corresponds to a vector that contains the weights of a given neuron in a layer of a multilayer neural network. Thus, generating the most diverse population of antibodies in $\Re^L$ corresponds to producing a set of neurons with well-distributed weight vectors. This way, the SAND approach will have to be applied separately to each layer of the network, as far as this layer contains more than a single vector. Another important aspect of the strategy is that, as we are generating vectors with unitary norms, these vectors can be normalized to force the activation of each neuron to occur near to the linear part of the activation functions, in order to avoid saturation.

4. Algorithms and Benchmarks

We compare the performance of SAND with four other methods to generate the initial set of weights: BOERS [2], WIDROW [16], KIM [11], OLS [13], and INIT [4]. All the five methods are applied to seven benchmark problems.

To specify the benchmark problems used, let *N* be the number of samples, SSE the desired sum squared-error (stopping criterion) and *net* the net architecture represented by [n_i-n_h-n_o]. Where n_i is the number of inputs, n_h is the number of hidden units and n_o is the number of outputs of the network

The benchmarks used for comparison were:

- parity 2 (XOR): $N = 4$, *net*: [2-2-1], SSE = 0.01;
- parity 3: $N = 8$, *net*: [3-3-1], SSE = 0.01;
- sin(*x*).cos(2*x*): $N = 25$, *net*:[1-10-1], SSE = 0.01;
- ESP: real-world problem used by [1]; $N = 75$, *net*: [3-10-5], SSE = 0.1;
- SOYA: another real-world problem used by [5], $N = 116$, *net*: [36-10-1], SSE = 0.1;
- IRIS: this benchmark is part of the machine learning database and is available in [15]; $N = 150$, *net*: [4-10-3], SSE = 0.15; and
- ENC/DEC: the family of encoder/decoder problem is very popular and is described in [6]. $N = 10$, *net*: [10-7-10].

The training algorithm used in all cases was the Moller scaled conjugate gradient [14], with the exact calculation of the second order information [17].

For each method and each benchmark problem we performed 10 runs. The results presented in Figure 2 correspond to the percentage of times each method produced the best performance in terms of maximum, minimum, mean and standard deviation of the number of epochs necessary for convergence to high quality solutions. This picture shows that SAND, INIT and OLS are superior to the others (BOERS, WIDROW, and KIM). If the name of a method do not appear in the comparison, it is because the approach does not converge given the maximum number of epochs, or converges to a poor local optimum. The advantage of SAND is that it does not make use of the training data to estimate the initial set of weights, like INIT and OLS.

5. Discussion

The proposed strategy (SAND) is inspired in the diversity preserving characteristic of the immune system. The SAND performance shows that neurons with well-distributed weight vectors leads to faster convergence rates. The necessity to re-scale the weight vectors reinforces the theory proposed by de Castro & Von Zuben [4]: initializing the weights in the approximately linear part of the neurons' activation function reduces numerical instabilities and results in improved convergence rates.

The performance of the proposed algorithm leads to two important conclusions concerning feedforward neural network initialization:

- It is necessary to avoid an initial set of weights that guides to the saturation of the neuron's response, what can be achieved by properly setting the weights' initial interval; and
- The generation of weight vectors mostly spread over the search-space results in smaller training times.

The method proposed also shows that there are still many biological phenomena in which to search for mechanisms and inspiration to solve computational intelligence problems, like neural network initialization, architecture optimization, learning, among others.

It is also important to stress that the proposed strategy does not take into account the training data, preparing the initial set of weights to deal appropriately with any input data, a process similar to the definition of the initial antibody repertoire of immune systems. In addition, the other approaches that were competitive, OLS and INIT, require the determination of inverse, or pseudo-inverse, matrices, being subject to numerical instabilities in cases the training samples contain linearly dependent vectors, i.e., redundancy, what is usually the case when training artificial neural networks.

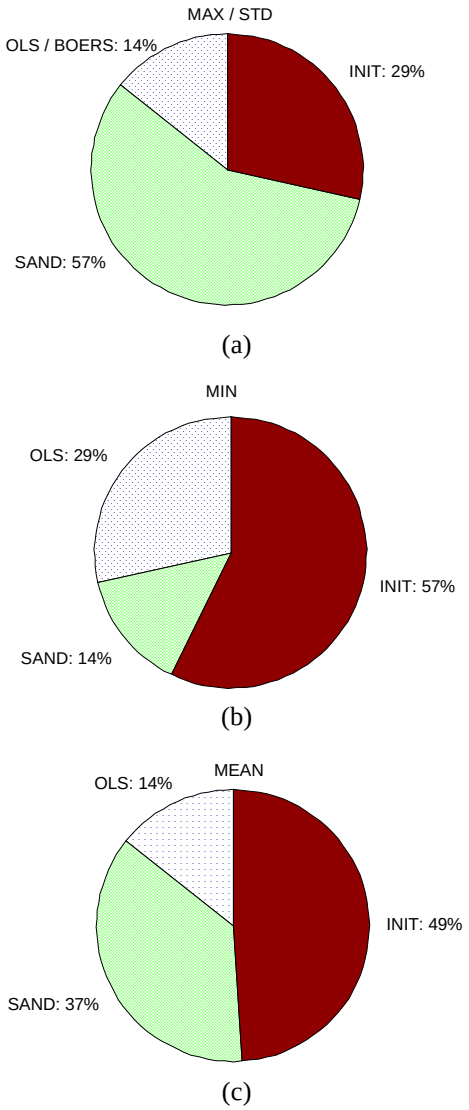

Figure 2: Performance comparison of the methods (considering 10 runs of 7 benchmark problems). (a) Percentage of times a method required the smallest maximum number of epochs for convergence and had the smallest standard deviation. (b) Percentage of times a method needed the smallest number of epochs for convergence. (c) Percentage of times a method presented the smallest mean number of epochs for convergence.

Acknowledgements

Leandro Nunes de Castro would like to thank FAPESP (Proc. n. 98/11333-9) and Fernando Von Zuben would like to thank FAPESP (Proc. n. 98/09939-6) and CNPq (Proc. n. 300910/96-7) for their financial support.

References

[1] Barreiros, J. A. L., Ribeiro, R. R. P., Affonso, C. M. & Santos, E. P., "Estabilizador de Sistemas de Potência Adaptativo com Pré-Programação de Parâmetros e Rede Neural Artificial", *Third Latin-American Congress: Eletricity generation and transmission*, pp.538-542, 1997.

[2] Boers, E. G. W. & Kuiper, H., "Biological Metaphors and the Design of Modular Artificial Neural Networks", Master Thesis, Leiden University, Leiden, Netherlands, 1992.

[3] de Castro, L. N., & Von Zuben, F. J., "Artificial Immune Systems: Part I – Basic Theory and Applications", Tech. Report RT–DCA 01/99, 1999.

[4] de Castro, L. N. & Von Zuben F. J., "A Hybrid Paradigm for Weight Initialization in Supervised Feedforward Neural Network Learning", *Proc. of the ICS'98*, Workshop on Artificial Intelligence, pp. 30-37, Taipei/Taiwan, R.O.C., 1998.

[5] de Castro, L. N., Von Zuben, F. J. & Martins, W., "Hybrid and Constructive Neural Networks Applied to a Prediction Problem in Agriculture". *Proc. of the IJCNN'98*, vol. 3, pp. 1932-1936, 1998.

[6] Fahlman, S. E., "An Empirical Study of Learning Speed in Back-Propagation Networks", *Tech. Rep.*, CMU-CS-88-162, School of Computer Science, Carnegie Mellon University, Pittsburg, PA, 1988.

[7] Haykin S., *Neural Networks – A Comprehensive Foundation*, Prentice Hall, 2nd Ed, 1999.

[8] Hertz, J., Krogh, A. & Palmer, R.G., *Introduction to the Theory of Neural Computation*. Addison-Wesley Publishing Company, 1991.

[9] Janeway Jr., C. A. & P. Travers, *Immunobiology The Immune System in Health and Disease*, Artes Médicas (in Portuguese), 2nd Ed, 1997.

[10] Kirkpatrick, S., Gelatt Jr., C. D. & Vecchi, M. P., "Optimization by Simulated Annealing", *Science, 220(4598)*, 671-680, 1987.

[11] Kim, Y. K. & Ra, J. B., "Weight Value Initialization for Improving Training Speed in the Backpropagation Network", *Proc. of IJCNN'91*, vol. 3, pp. 2396-2401, 1991.

[12] Kolen, J. F. & Pollack, J. B., "Back Propagation is Sensitive to Initial Conditions", *Technical Report TR 90-JK-BPSIC*, 1990.

[13] Lehtokangas, M., Saarinen, J., Kaski, K. & Huuhtanen, P., "Initializing Weights of a Multilayer Perceptron by Using the Orthogonal Least Squares Algorithm", *Neural Computation*, vol. 7, pp. 982-999, 1995.

[14] Moller, M. F., "A Scaled Conjugate Gradient Algorithm for Fast Supervised Learning", *Neural Networks*, vol. 6, pp. 525-533, 1993.

[15] ftp://ftp.ics.uci.edu/pub/machine-leraning-databases

[16] Nguyen, D. & Widrow, B., "Improving the Learning Speed of two-layer Neural Networks by Choosing Initial Values of the Adaptive Weights", *Proc. IJCNN'90*, vol. 3, pp. 21-26, 1990.

[17] Pearlmutter, B. A., "Fast Exact Calculation by the Hessian", *Neurocom*, vol. 6, pp. 147-160, 1994.

[18] Perelson, A. S. & Oster, G. F., "Theoretical Studies of Clonal Selection: Minimal Antibody Repertoire Size and Reliability of Self-Nonself Discrimination", *J. theor. Biol., 81*, 645-670, 1979.

[19] Shepherd, A. J., *Second-Order Methods for Neural Networks – Fast and Reliable Methods for Multi-Layer Perceptrons*, Springer, 1997.

[20] Smith, D. J., Forrest, S., Hightower, R. R. & Perelson, S. A., "Deriving Shape Space Parameters from Immunological Data", *J. theor. Biol.*, 189, pp. 141-150, 1997.