Neural Processing Letters **7:** 1–4, 1998.

Characterization of the Sonar Signals Benchmark

JUAN MANUEL TORRES MORENO AND MIRTA B. GORDON⋆
Département de Recherche Fondamentale sur la Matière Condensée, CEA/Grenoble – 17, Av. des Martyrs – 38054 Grenoble Cédex 9. France
E-mail: Mirta.Gordon@cea.fr

Key words: classification problems, generalization, learning, perceptrons, sonar targets

Abstract. We study the classification of sonar targets first introduced by Gorman & Sejnowski (1988). We discovered that not only the training set *and* the test set of this benchmark are both linearly separable, although by different hyperplanes, but that the *complete* set of patterns, training and test patterns together, is also linearly separable. The distances of the patterns to the separating hyperplane determined by learning with the training set alone, and to the one determined by learning the complete data set, are presented.

It has become a current practice to test the performance of learning algorithms on realistic benchmark problems. The underlying difficulty of such tests is that in general these problems are not well characterized: given a solution to the classification problem, it is impossible to decide whether a better one exists.

The sonar signals benchmark [1] has been widely used to test learning algorithms [2–10]. In this problem the classifier has to discriminate if a given sonar return was produced by a metal cylinder or by a cylindrically shaped rock in the same environment. The benchmark contains 208 preprocessed sonar spectra defined by $N = 60$ real values in the range [0, 1], and their corresponding class. Among these, the first $P = 104$ patterns are usually used as the *training set* to determine the classifier parameters. The fraction of misclassified patterns among the remaining $G = 104$ spectra, the *test set*, is used to estimate the generalization error produced by the learning algorithm.

We studied this benchmark with Minimerror, a training algorithm for *binary* perceptrons [11, 12] that allows for a gradient search of normalized weights $\vec{w}$, $\vec{w} \cdot \vec{w} = N$, through the minimization of a parameterized cost function,

$$E = \frac{1}{2} \sum_{\mu=1}^{P} V\left(\frac{\tau^{\mu} \vec{w} \cdot \vec{\xi}^{\mu}}{2T\sqrt{N}} \right), \tag{1}$$

$$V(x) = 1 - \tanh(x). \tag{2}$$

⋆ Also at Centre National de la Recherche Scientifique (CNRS); author to whom correspondence should be sent.

where $\vec{\xi}^{\mu}$ is the input pattern ($\mu = 1, \cdots, P$), $\tau^{\mu} = \pm 1$ its class. We arbitrarily defined $\tau = +1$ for mines and $\tau = -1$ for rocks. The parameter T, called temperature (for reasons related to the interpretation of the cost function), defines an effective window width on both sides of the separating hyperplane. The derivative $dV(x)/dx$ is vanishingly small outside this window. Therefore, if the minimum of cost (1) is searched through a gradient descent, only the patterns at a distance $d^{\mu} \equiv |\vec{w} \cdot \vec{\xi}^{\mu}|/\sqrt{N} < 2T$ will contribute significantly to learning. The algorithm Minimerror implements this minimization starting at high temperature. The weights are initialized with Hebb's rule, which is the minimum of (1) in the high temperature limit. Then, T is slowly decreased upon the successive iterations of the gradient descent – a procedure called *deterministic annealing* – so that only the patterns within the narrowing window of width 2T are effectively taken into account to calculate the correction $\delta\vec{w} = -\epsilon\, \partial E/\partial\vec{w}$ at each time step, where ϵ is the learning rate. Thus, the search of the hyperplane becomes more and more local as the number of iterations increases. In practical implementations, it was found that convergence is considerably speeded-up if already learned patterns are considered at a lower temperature T_L than not learned ones, $T_L < T$. The algorithm Minimerror has three free parameters: the learning rate ϵ of the gradient descent, the temperature ratio T_L/T, and the annealing rate δT at which the temperature is decreased. At convergence, a last minimization with $T_L = T$ is performed. Further details of the implementation of Minimerror may be found in [11, 12].

Coming back to the sonar signals, we found that not only both the training set (*i.e.* the first $P = 104$ patterns hereafter called the 'standard' training set) and the test set (*i.e.* the last $G = 104$ patterns) of the benchmark are linearly separable, a fact already reported [13, 14], but that also the complete set of $P+G = 208$ patterns is linearly separable. The algorithm Minimerror finds the separating hyperplanes within a broad range of parameter values. The generalization error of the weights $\vec{w}_P$ that separate the standard training set is $\epsilon_g \cong 22\%$, corresponding to 23 classification errors on the test set. A lower generalization error may be obtained through early stopping, *i.e.* by stopping the algorithm before convergence. Our best generalization performance, $\epsilon_g \cong 15\%$ (16 errors), was obtained by stopping with 8 training errors (we denote $\vec{w}_{Pe}$ the corresponding weights). However, the overall performance (training and test errors added together) is worse than the one obtained with the weights $\vec{w}_P$. By training with the patterns usually used as test set (*i.e.* the last $G = 104$ patterns of the sonar data base) we determined weights $\vec{w}_G$ that linearly separate the test set. The corresponding generalization error estimated using the P first patterns as test set is $\epsilon_g \cong 23\%$ (24 errors). Finally, by training with the complete set of $P + G = 208$ patterns, weights $\vec{w}_{P+G}$ separating *all* the patterns could be found, showing that this benchmark is linearly separable.

The weights $\vec{w}$ obtained by training with the different sets are normal to the corresponding separating hyperplanes. The projections of the patterns onto the unitary vectors $\vec{w}/\sqrt{N}$, $d^{\mu} \equiv \vec{w} \cdot \vec{\xi}^{\mu}/\sqrt{N}$, are proportional to the weighted sum; $|d^{\mu}|$ is the distance of pattern μ to the separating hyperplane, whereas $\mathrm{sign}(d^{\mu})$ is

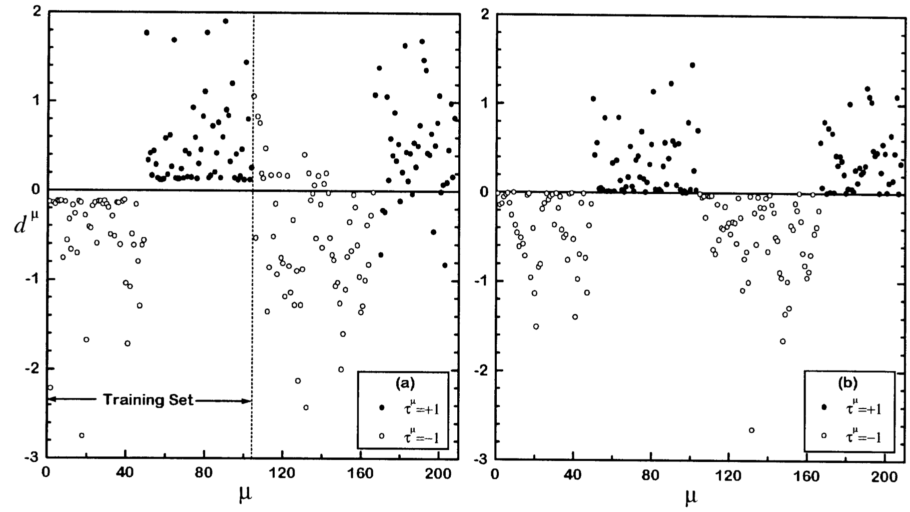

Figure 1. Distance of the patterns to the separating hyperplane, with a sign corresponding to the actual perceptron's output. The correct class is τ^μ ($\tau^\mu = +1$ for mines, $\tau^\mu = -1$ for rocks). (a) Hyperplane determined with the standard training set (contains the first $P = 104$ patterns of the sonar data set), showing the 23 errors on the test set. (b) Hyperplane determined with the complete sonar data set of $P + G = 208$ patterns.

the actual network's output to pattern μ. We represented on figure 1 the values of d^μ corresponding to $\vec{w}_P$ and $\vec{w}_{P+G}$, as a function of the pattern number. It may be seen that weights $\vec{w}_P$ correspond to a robust solution: there is a gap, of width $\kappa = 0.1226$ free of training patterns, on both sides of the hyperplane. This gap is much more narrow ($\kappa = 0.00284$) – hardly visible on the figure – for the solution separating the complete data set, showing that this is a much harder problem.

Finally, let us point out that the different weights are *not* close to each other, as may be seen by pairwise comparison of the overlaps $R_{a,b} = \vec{w}_a \cdot \vec{w}_b / N$, that should be 1 for identical solutions. The overlaps between our different solutions are $R_{P,G} = -0.124$, $R_{P,P+G} = 0.580$, $R_{G,P+G} = 0.543$; whereas the corresponding overlaps with early stopping results obtained with the P patterns of the standard training set are $R_{Pe,P} = 0.516$, $R_{Pe,G} = 0.345$, $R_{Pe,P+G} = 0.525$. These results are not surprising, as it is well known that typically, *i.e.* with probability close to 1, up to $2N$ not correlated patterns are linearly separable in N dimensions [15], and this number increases if patterns are correlated [16].

Acknowledgment

The autor J. Manuel Torres Moreno has been supported by CONACYT and UAM-Azcapotzalco (México), grant 65659.

References

1. 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.
2. M. Berthold, “A probabilistic extension for the DDA algorithm”, in IEEE International Conference on Neural Networks, pp. 341–346, Washington, 1996.
3. M.R. Berthold and J. Diamond, “Boosting the performance of RBF networks with dynamic decay adjustment”, in G. Tesauro, D. Touretzky, and T. Leen, editors, Advances in Neural Information Processing Systems, Vol. 7, pp. 521–528, The MIT Press, 1995.
4. J. Bruske and G. Sommer, “Dynamic cell structures”, in G. Tesauro, D. Touretzky, and T. Leen, editors, Advances in Neural Information Processing Systems, Vol. 7, pp. 497–504, The MIT Press, 1995.
5. B. Chakraborty and Y. Sawada, “Fractal connection structure: Effect on generalization supervised feed-forward networks”, in IEEE International Conference on Neural Networks, pp. 264–269, Washington, 1996.
6. M. Karouia, R. Lengellé and T. Denoeux, “Performance analysis of a MLP weight initialization algorithm”, in Michel Verleysen (ed.) European Symposium on Artificial Neural Networks, pp. 347–352, Brussels, 1995, D facto.
7. A. Roy, S. Govil and R. Miranda, “An algorithm to generate radial basis function (rbf)-like nets for classification problems”, Neural Networks, 8(2): 179–201, 1995.
8. A. Roy, L. Kim and S. Mukhopadhyay, “A polynomial time algorithm for the construction and training of a class of multilayer perceptron”, Neural Networks, 6(1): 535–545, 1993.
9. Y. Shang and B.W. Wha, “A global optimization method for neural networks training”, in IEEE International Conference on Neural Networks, pp. 7–11, Washington, 1996.
10. Brijesh K. Verma and Jan J. Mulawka, “A new algorithm for feedforward neural networks”, in Michel Verleysen (ed.) European Symposium on Artificial Neural Networks, pp. 359–364, Brussels, 1995, D facto.
11. M.B. Gordon and D. Berchier, “Minimerror: A perceptron learning rule that finds the optimal weights”, in Michel Verleysen, editor, European Symposium on Artificial Neural Networks, pp. 105–110, Brussels, 1993. D facto.
12. B. Raffin and M.B. Gordon, “Learning and generalization with minimerror, a temperature dependent learning algorithm”, Neural Computation, 7(6): 1206–1224, 1995.
13. M. Hoehfeld and S. Fahlman, “Learning with limited numerical precision using the cascade correlation algorithm”, Technical Report CMU-CS-91-130, Carnegie Mellon University, 1991.
14. J.-M. Torres Moreno and M. Gordon, “An evolutive architecture coupled with optimal perceptron learning for classification”, in Michel Verleysen (ed.) European Symposium on Artificial Neural Networks, pp. 365–370, Brussels, 1995, D facto.
15. T.M. Cover, “Geometrical and statistical properties of systems of linear inequalities with applications in pattern recognition”, IEEE Transactions on Electronic Computers, EC–14: 326–334, 1965.
16. E. Gardner, “Maximum storage capacity in neural networks”, Europhysics Letters, 4: 481–485, 1987.