ESANN'1997 proceedings - European Symposium on Artificial Neural Networks
Bruges (Belgium), 16-17-18 April 1997, D-Facto public., ISBN 2-9600049-7-3, pp. 115-120

Knowledge extraction from neural networks for signal interpretation

F. Alexandre and J.-F. Remm
CRIN-INRIA BP 239
F-54506 Vandœuvre
falex@loria.fr

Abstract

Artificial neural networks have proved their ability to perform classification tasks. This ability is not satisfactory when expertise of the application domain is not available or when experts want to know more about hints that led to the decision. This leads presently to a great amount of work for knowledge or rule extraction from neural networks. In this paper, we propose a technique able to extract rules and to explain the functioning of the hidden layers of a multilayer perceptron. The first step consists in pruning the network with the classical OBD algorithm. Then, tightening of the sigmoidal transfer function can simply result in such knowledge extraction. This principle has been first tested on an application of signal interpretation in the radar domain.

1 Introduction

In spite of their ability to perform such difficult tasks as classification, function approximation, optimization or control, Artificial Neural Networks (ANNs) are often judged unsatisfactory, because they can give no explanation about their decision: ANNs are black boxes. Knowledge extraction from ANNs is thus a very important question today.

Automatic radar target identification has been rarely tackled by ANNs. This kind of work is all the more difficult as no expert is available to explain the meaning of the signal. It is thus an interesting application for neuronal knowledge extraction. We first introduce the application domain, explain why a knowledge extraction is pertinent here and explain our method on an example from the radar domain.

2 The radar domain

The principle of a radar [6], is to illuminate a target by transmitting properly designed signals, and to receive and process the returned echoes. Range localisation is obtained by measuring the time delay between the transmitted and received signals. A promising way to identify radar targets is to use a radar transmitting short pulses in order to obtain radar resolution cells thinner than the target spatial extent. The target is no more seen as a point but composed of scattering centres [2].

We first tested neuronal techniques on automatic stationary target identification for high range resolution radars. This study was realized on real targets, namely three kinds of terrestrial vehicle named target C, D, E. The data consist in range profiles of each target for different aspect angles.

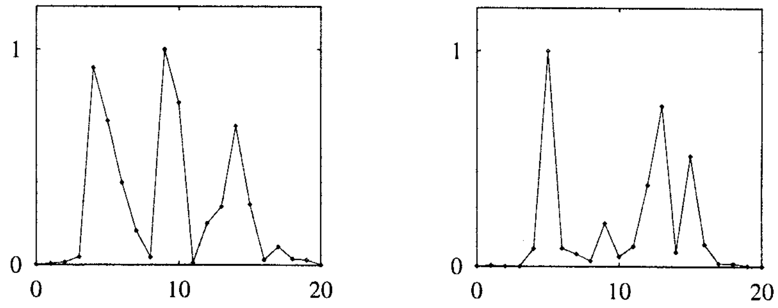

Figure 1: Examples of range profile

Fig. 1 gives an example of such signatures for different kinds of target. The y axis stands for the energy; the x axis represents the sampling along the radial distance. It must be underlined that such signals are very noisy and that two close signatures can be very different, because scattering centres can disappear or loose a part of their energy.

Comparison with other classical identification techniques has been discussed in [4]. We report here trials carried out with Multi Layer Perceptrons (MLPs) with the backpropagation learning algorithm, a momentum and a learning coefficient per weight. We chose, for all the experiments reported below, an architecture with 20 neurons on the input layer (representing the 20 points of sampling on the radial distance), 7 for the hidden layer and 3 for the output layer (representing the 3 different possible targets).

The first problem that we faced was the one of preprocessing. One solution was to choose the one that is used with classical methods. In this case, the input of the MLP is the auto-correlation of the input profile [4]. As auto-correlation techniques deeply modify the shape of the signal, we also decided to test this connectionist supervised method on the rough signal, without auto-correlation preprocessing. In this case, we had to design an efficient localisation of the beginning of the signal. Here this localisation the obtained through a correlation between the signal and a step.

Figure 2 below gives some results with the architecture and the corpus described above, and after 10,000 learning epochs. The y axis represents the recognition rate and the x axis the number of averaged measures for recognition. Results obtained with a classical bayesian approach are also reported, for comparison.

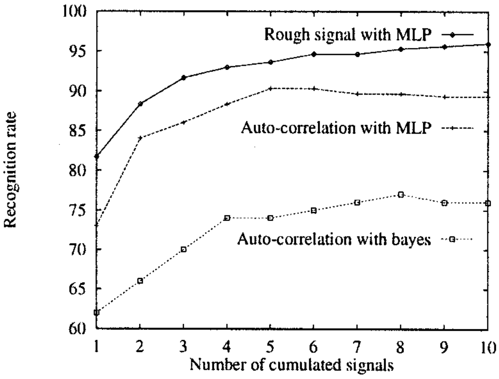

Figure 2: Recognition results

This figure indicates good results with neuronal techniques. It also indicates that auto-correlation seems not appropriate as a preprocessing. These results, though better with regard to other classical approaches are not sufficient for several reasons. First, the recognition rate itself is not high enough to directly include the MLP in an automatic identification system. Second, discussions with experts of the domain showed that they were not very interested in using, for such a task, a black box unable to explain its decision. Moreover, they were also very interested in having a tool able to help them to analyse and describe the signal. That is the reason why, on the basis of these results, we tried to imagine how the MLP could give an explanation of the hints, built during its learning and used for the classification.

3 Knowledge extraction from a MLP

Some authors have proposed to extract hints and rules from ANNs. These methods have constraints concerning the input data (boolean inputs in [3]) or the sigmoidal transfer function (not used on its whole spectrum in [5]) or with regard to the need for pre-existing knowledge (like in [7]).

Our method consists in learning the task with a classical MLP and then pruning this network with a classical algorithm like OBD [1]. From this pruned network, we get rid of the non linear aspect of the sigmoidal transfer function of the neurons while stiffening the sigmoid during new learning processes. Then, knowledge can be easily extracted. It is divided in three steps, each corresponding to a kind of knowledge: (i) analysis of the hidden layer (selection of pertinent hints), (ii) analysis of the input layer (giving a meaning to the hints), (iii) recombination of both (giving rules).

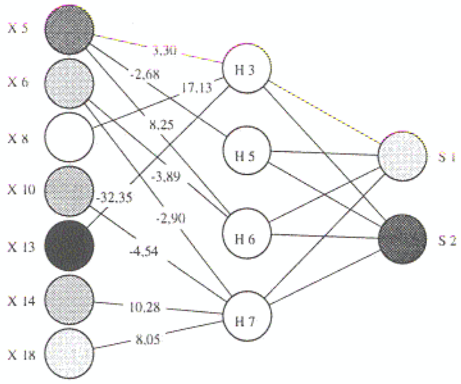

Figure 3: Simplified network at 10°

We illustrate it with an example from the radar domain. This example is concerned with knowledge extraction from a network differentiating targets C and E at 10°. The fully connected network has an accuracy of 87.5% on a test corpus. The network reported in figure 3 is obtained after pruning and yields 85.5% of recognition rate, even after the sigmoid stiffening.

Analysis of the hidden layer The interest of sigmoid stiffening is that each hidden neuron can only have two values. Thus, if n is the number of hidden neurons, there remains 2^n possible configurations of the hidden layer and all of them are not activated by the corpus.

In our example, on the 16 possible vectors, 8 are used :

HL_i	$(H_3, H5, H6, H7)$	target
HL_1	$(-1, -1, -1, -1)$	C
HL_2	$(-1, -1, -1, 1)$	C
HL_3	$(-1, -1, 1, -1)$	C
HL_4	$(-1, -1, 1, 1)$	C
HL_5	$(1, -1, -1, 1)$	C
HL_6	$(1, -1, 1, 1)$	C
HL_7	$(1, -1, -1, -1)$	E
HL_8	$(1, -1, 1, -1)$	E

The boolean simplification of the expressions gives :

$$S_1 = \text{Target C} = \overline{H_3}\,\overline{H_5} + H_3\,\overline{H_5}\,H_7$$

$$S_2 = \text{Target E} = H_3\,\overline{H_5}\,\overline{H_7}$$

Analysis of the input layer We now turn to discriminant neurons of hidden layer and to neurons of the input layer that are still connected to the hidden layer, after the pruning process. A discriminant neuron is a neuron that appears in expressions of discriminant domains. In our example, discriminant neurons are neurons H_3, H_5 and H_7 whereas H_6 is not a discriminant neuron. If k neurons of the input layer are connected to the discriminant neuron H_i , these input neurons define an hyperplane whose equation is given by the network.

$$\begin{aligned}
H_3 &= 3.30X_5 + 17.13X_8 - 32.35X_{13} > 0 \\
H_5 &= -2.68X_5 > 0 \\
H_6 &= 8.25X_5 - 3.89X_6 > 0 \\
H_7 &= -2.90X_6 - 4.54X_{10} + 10.28X_{14} + 8.05X_{18} > 0
\end{aligned}$$

We thus extract the following rules:

If $(3.30X_5 + 17.13X_8 - 32.35X_{13} < 0)$
And $(2.68X_5 > 0)$ Then Target C

If $(3.30X_5 + 17.13X_8 - 32.35X_{13} > 0)$
And $(2.68X_5 > 0)$
And $(-2.90X_6 - 4.54X_{10} + 10.28X_{14} + 8.05X_{18} > 0)$ Then Target C

If $(3.30X_5 + 17.13X_8 - 32.35X_{13} > 0)$
And $(2.68X_5 > 0)$
And $(-2.90X_6 - 4.54X_{10} + 10.28X_{14} + 8.05X_{18} < 0)$ Then Target E

Adaptation of the algorithm The boolean simplification can be slightly modified:

- If some neurons always vary in the same direction, whatever vector HL_i, it is possible to eliminate this input from discriminant neurons. In our example, it is the case with H_5.
- If, whatever vector HL_i, we always have $H_j = H_k$ or $H_j = \overline{H_k}$, then it is possible to keep, as a discriminant vector either H_j or H_k and reject the other.

In our example, this gives the following rules:

$$\text{Target C} = \overline{H_3} + H_3\,H_7$$

$$\text{Target E} = H_3\,\overline{H_7}$$

4 Conclusion

The knowledge extraction method that we have presented here proved its efficiency on several examples of radar target interpretation. Some were more simple (pruning can sometimes remove a very large part of the network). Other were more difficult and in this case the help of experts was mandatory. Extracted rules and hints provided radar experts with many new data. It is worth noticing here that hints generally seem more interesting to them, since rules give too complicated numerical relations, like hyperplan equations. This is rather easy to understand. This kind of approach corresponds to a need for a cognitive interpretation of a phenomenon. Thus, the emergence of a subsymbolic representation is highly preferable to a numerical representation which is closer to the black box effect of classical ANNs. Today, the pruned ANN is used to give an explanation of the choice, but the entire ANN is also used for classification. Deeply pruning the ANN to get much more simple rules won't decrease the recognition performances of the whole system, since the entire ANN is still present for this aspect. This work must also be understood as a contribution to ANN knowledge extraction. We are now testing this algorithm on other applications and try to propose a general methodology for such an approach.

The authors wish to thank the Direction des Recherches Etudes et Techniques (DRET) which gave them the opportunity to treat this subject and allowed them to publish this text.

References

[1] Y. Le Cun, J.S. Denker, and S.A. Solla. Optimal brain damage. In *Advances in Neural Information Processing Systems II (Denver 1989)*, 1990.

[2] H.J. Li and S.H Yang. Using range profile as feature vectors to identify aerospace objects. In *IEEE Transactions on antennas and propagation*, pages 261–268, 1993.

[3] E. Po, R. Hayward, and J. Diederich. Ruleneg : extracting rules from a trianed ann by stepwise negation. Technical report, QUT NRC, 1994.

[4] J.-F. Remm, F. Alexandre, and L. Savy. Automatic Radar Target Identification using Neural Networks. In *Proceedings International Conference of Artificial Neural Nets and Genetic Algorithms*, 1995.

[5] R. Setiono and H. Liu. Symbolic representation of neural networks. *Computer*, 29(3):71–77, 1996.

[6] M.I. Skolnik. *Introduction to radar systems.* Mc Graw-Hill, New-York, 1980.

[7] G.G. Towell and J.W.Shavlik. Extracting refined rules from knowledge-based neural networks. *Machine Learning*, 13(1):71–101, Oct. 1993.