ESANN'1995 proceedings - European Symposium on Artificial Neural Networks
Brussels (Belgium), 19-20-21 April 1995, D-Facto public., ISBN 2-9600049-3-0, pp. 259-264

Derivation of a New Criterion Function based on an Information Measure for Improving Piecewise Linear Separation Incremental Algorithms

J. Cugueró, J. Madrenas, J. M. Moreno, J. Cabestany

Departament d'Enginyeria Electrònica
Universitat Politècnica de Catalunya (Spain)

Abstract. When a single neuron is trained in a PLS algorithm stage, it is desirable that it captures as much information as possible from the training set under consideration. There exist several criterion functions that intend to measure this information gain. In this paper, a criterion function is developed from very simple assumptions. The criterion is further simplified to allow easy implementation and tested to show the improvement it yields.

1. Introduction

From the very first classical neural algorithms there has been the tendency to reduce the a priori constraints on the network architecture allowing the training process to get this information from the training set. Evolutive algorithms are the kind of algorithms that allow this architecture-tunning to the problem.

Among incremental evolutive algorithms, Piecewise Linear Separation (PLS) algorithms, which are mainly devoted to classification tasks, are very popular. The criterion derived will be tested using the PLS Neural Trees [1] algorithm. Given a training set with two different pattern classes, Neural Trees algorithm looks for a neuron that either divides completely the two classes, or splits the initial training set into two smaller sets so that, recursively, two new neurons can be trained on them. So a binary tree grows from the first neuron until a stopping criterion is met. One possible criterion can be to generate units until the whole training set is correctly classified.

Single neuron training

Perceptron [2] or Pocket [3] are the usual training algorithms for the neurons generated by PLS algorithms. The Pocket algorithm is a variation of the Perceptron algorithm that consists on testing the neuron with the whole training set everytime neuron synapses change. The neuron is kept in a "pocket" if it results to be

ESANN'1995 proceedings - European Symposium on Artificial Neural Networks
Brussels (Belgium), 19-20-21 April 1995, D-Facto public., ISBN 2-9600049-3-0, pp. 259-264

better than the previous ones. There is a variety of criteria to measure the quality of a neuron some of which are shown in [4][5][6].

2. Derivation of an information criterion

In order to derive an information criterion, every neuron encountered during training will be assigned an a priori probability measure. Minimization of this probability measure will be the best neuron selecting criterion. Information Theory provides a justification of this selection: the less probable an event is, the more information its occurrence gives. Thus, neurons less probable a priori obtained during training will be the ones that give more information about the training set.

Suppose a two class training set with *clas0* patterns belonging to class 0 and *clas1* patterns belonging to class 1. This training set will be one of the

$$totalDichots - C_{clas0+clas1}^{clas0} \tag{1}$$

possible dichotomies. Where C is a binomial coefficient. Since no a priori information about the dichotomies is available they will be assumed equiprobable using Laplace principle.

Suppose that a neuron **n** encountered during training performs the following separation of the training set: in its 0 side (the side where the neuron decides that a pattern belongs to class 0) there are *side0_clas0* patterns belonging to class 0 and *side0_clas1* patterns belonging to class 1; and in its 1 side there are *side1_clas1* patterns of class 1 and *side1_clas0* patterns of class 0. The total number of possible dichotomies compatible with neuron **n** is

$$dichots(n) - C_{side0\_clas0+side0\_clas1}^{side0\_clas0} \; C_{side1\_clas1+side1\_clas0}^{side1\_clas1} \tag{2}$$

Thus, it is possible to assign a probability measure to neuron **n** since it selects a fraction of the total possible and equiprobable dichotomies of the training set. The following quotient expresses this probability:

$$p(n) - \frac{dichots(n)}{totalDichots} \tag{3}$$

ESANN'1995 proceedings - European Symposium on Artificial Neural Networks
Brussels (Belgium), 19-20-21 April 1995, D-Facto public., ISBN 2-9600049-3-0, pp. 259-264

This is the criterion Pocket algorithm has to minimize.

If we want to give an Information theoretical interpretation to the previous measure we must express it as an information quantity I(**n**) = -log p(**n**),

$$I(\boldsymbol{n}) - \log C_{clas0+clas1}^{\;clas0} - \log C_{side0\_clas0+side0\_clas1}^{\;side0\_clas0} - \log C_{side1\_clas1+side1\_clas0}^{\;side1\_clas1} \tag{4}$$

The criterion expressed as an information quantity needs to be maximized by the Pocket algorithm. Equation (5) criterion is an equivalent representation of the Information criterion shown by Nadal and Toulouse [7].

The high computational cost implicit in the evaluation of equations (4) or (5) suggests the search of a simpler criterion that shares their crucial properties.

3. Derivation of a simpler criterion

Two intuitive arguments will be used to derive a simpler criterion:

- A neuron gives the more information about class 1 (or about class 0) the larger it is the difference between the memorization of the patterns in the 1 side (or the 0 side) of that neuron and the a priori (without any neuron) memorization of class 1 (or class 0) patterns.

- A neuron gives the more information about class 1 (or about class 0) the more patterns are in the 1 side (or the 0 side) of that neuron.

The previous insights allow the construction of two different criteria, one for each class of patterns. For class 1 the criterion will be:

$$\tilde{I}_{class1}(\boldsymbol{n}) - (side1\_clas1 + side1\_clas0) \, | m_{side\,1} - m_{a\,priori\,class1} | \tag{5}$$

where memorizations $m_{side\ 1}$ and $m_{a\ priori\ class\ 1}$ are

$$m_{side\,1} - \frac{side1\_clas1}{side1\_clas1 + side1\_clas0}$$
$$m_{a\,priori\,class1} - \frac{clas1}{clas1 + clas0} \tag{6}$$

After simplifying we obtain

$$\tilde{I}_{class1}(n) = \left| \frac{clas0\ side1\_clas1 - clas1\ side1\_clas0}{clas0+clas1} \right| \tag{7}$$

A similar reasoning for class 0 yields:

$$\tilde{I}_{class0}(n) = \left| \frac{clas0\ side0\_clas1 - clas1\ side0\_clas0}{clas0+clas1} \right| \tag{8}$$

It can be proved that criteria (8) and (9) are different representations of exactly the same one; this can be seen by substituting *side0_clas1* with its equivalent expression *clas1 - side1_clas1* and *side1_clas0* with *clas0 - side0_clas0.*
Since the criterion will only be evaluated for maximization, there is no need of preserving denominators of (8) or (9) because they do not depend on the particular neuron **n** under consideration. Thus, two different expressions for the same criterion can be used

$$I'(n) = |clas0\ side1\_clas1 - clas1\ side1\_clas0| = |clas0\ side0\_clas1 - clas1\ side0\_clas \tag{9}$$

The above I'(**n**) criterion like the I(**n**) information criterion in (5) is maximum when neuron **n** separes linearly the two classes and is minimum (equal to zero) when neuron **n** does not split the set into two smaller subsets.

4. Test of the I' criterion

The Neural Trees algorithm and two artificial databases (clouds and gauss_2D) will be used to test the I' criterion. Both databases are two-class, two-dimensional and have 5000 patterns: 2500 each class. They are composed of gaussian distributions. A full description appears in [8].

Single neurons will be trained using the Pocket algorithm with a threshold t for the I' criterion, so that a neuron **n** will not be accepted if I'(**n**)<t. The threshold t is a simple way to control the number of units of the final network: the higher the threshold t, the lower the number of units of the network. When a training process on a training subset fails to find a neuron, the class of the patterns in that zone of the domain will be decided by the previous neuron in the binary tree hierarchy based on the ratio of patterns of each class in the training subset.

Various t values will be considered in order to obtain different network sizes. For every t value the training set will be divided in ten parts nine of which will be used as the training set and the tenth will be used as the test set to measure the generalization ability of the network. Ten networks will be trained with the ten possible training and test sets. The mean values for the number of neurons and the generalization have been calculated.

In order to compare, the same process is repeated maximizing the following criterion proposed in [9]:

$$J(n) = \frac{side0\_clas0}{1+side0\_clas1} + \frac{side1\_clas1}{1+side1\_clas0} \qquad (10)$$

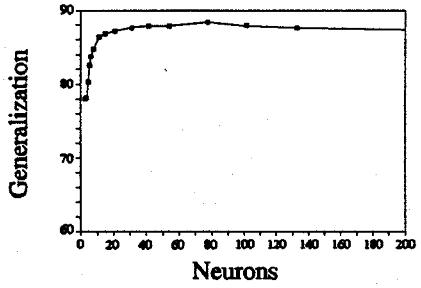

Fig. 1. I' criterion with clouds

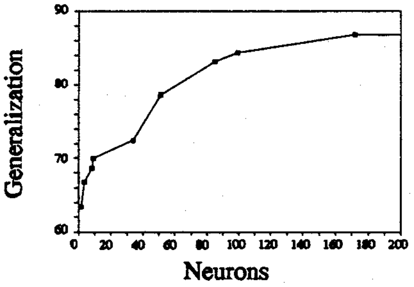

Fig. 2. J criterion with clouds

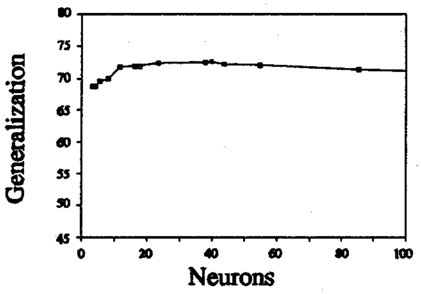

Fig. 3. I' criterion with gauss_2D

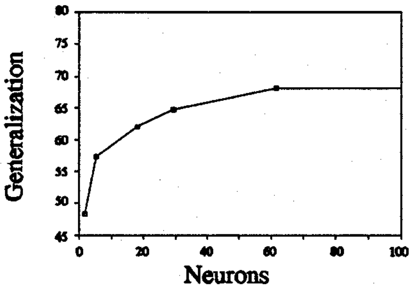

Fig. 4. J criterion with gauss_2D

Fig. 1 and Fig. 2 show the generalization ability as a function of the network size for I' and J criteria and the clouds database. Pocket has run 500 iterations for I' and 1000 iterations for J. More neurons are needed to achieve the same generalization when the J criterion is used. For example, I' criterion needs a mean of 7.6 neurons to achieve a 84.74% mean generalization while J criterion needs 99.4 neurons to achieve approximately the same generalization.

Fig. 3 and Fig. 4 show the same plot for the gauss_2D database and the same number of iterations as with clouds. The plots show again the superior performance of the I' criterion. For example, the I' criterion achieves a 68.72% mean generalization with a mean of 3.3 neurons while the J criterion needs a mean of 61.70 units for approximately the same generalization.

5. Conclusions

An information criterion to measure the goodness of a neuron during a training process has been derived from few simple hipothesis. An alternative criterion computationally simpler that preserves maximum and minimum values of the former has been obtained. The new criterion has been tested and compared with another one yielding better results in terms of units to achieve a certain level of generalization. That performance is due to the compression implicit in any information criterion. Furthermore, the simplicity of the criterion makes it especially efficient in CPU time or in hardware implementations.

References

1. J. A. Sirat, J. P. Nadal: Neural Trees: A New Tool for Classification. Technical Report. Laboratoires d'Electronique Philips. (1990)
2. D. E. Rumelhart, J. L. McClelland: Parallel Distributed Processing: Explorations in the Microstructure of Cognition. MIT Press. (1986)
3. S. I. Gallant: Optimal Linear Discriminants. Proc. of the 8th. Intl. Conf. on Pattern Recognition, vol. 2, 849-854. Paris. (1986)
4. J. M. Moreno, F. Castillo, J. Cabestany: Enhanced Unit Training for Piecewise Linear Separation Incremental Algorithms. Proc. of ESANN 93, 33-38, Brussels. (1993)
5. J. M. Moreno, F. Castillo, J. Cabestany: Optimized Learning for Improving the Evolution of Piecewise Linear Separation Incremental Algorithms. New Trends in Neural Computation, 272-277, J. Mira, J. Cabestany, A. Prieto (eds.), Springer-Verlag. (1993)
6. J. M. Moreno, F. Castillo, J. Cabestany: Improving Piecewise Linear Separation Incremental Algorithms using Complexity Reduction Methods. Proc. of IWANN 93, 141-146, Sitges. (1993)
7. J. P. Nadal, G. Toulouse: Information Storage in Sparsely-Coded Memory Nets. Network, vol. 1, 61-74. (1990)
8. F. Blayo et al., ELENA ESPRIT BRA 6891 Report R2-B1-P Task B1: Databases. (1994)
9. S. Knerr, L. Personnaz, G. Dreyfus: A New Approach to the Design of Neural Classifiers and its Application to the Automatic Recognition of Handwritten Digits. Proc. of IJCNN 91, 91-96, Seattle. (1991)