New Neural Network Pruning and its application to sonar imagery

Pascal Galerne, Koffi Clément Yao, Gilles Burel

HAL Id: hal-03222606

https://hal.univ-brest.fr/hal-03222606

Submitted on 18 Mar 2023

New Neural Network Pruning and its application to sonar Imagery

P. Galerne(*), K. Yao(*), G. Burel(#)
(*) Ecole Navale, Lanvéoc Poulmic, 29240 Brest Naval, France
email : name@poseidon.ecole-navale.fr
phone : 33 (2) 98 23 38 47
fax : 33 (2) 98 23 38 57
(#) University of Brest, Département d'Electronique,
6, av. Le Gorgeu, BP 809, 29285 BREST cedex, France
email : Gilles.Burel@lest-gw.univ-brest.fr
phone : 33 (2) 98 01 62 46
fax : 33 (2) 98 01 63 95

ABSTRACT

In this paper, we propose a neural network approach for automatic classification of underwater objects on sonar images. A major problem with sonar imagery applications is the difficulty to obtain large databases for training. Real sonar devices are costly and staged experiments where objects are well known and manually placed are rare because of cost. We show that the simultaneous use of parameters extraction and neural network pruning can significantly help to obtain good generalization rates (despite the lack of large training databases) and to reduce the complexity of the classifier.

1. INTRODUCTION

The technological improvement of high frequency sonars allows to prospect widely sea-bottom areas with more accuracy. In counterpart, information stream scrutinized by an operator whose job is to detect and classify objects on sonar images has noticeably increased. Therefore, the exploitation of the collected data has to be achieved with an automatic processing chain.

The classification of an object lying on the seafloor is based on the analysis of its cast shadow shape. To achieve this task, the system detect each shadow contained in the image by a two class segmentation process (shadow class and reverberation class). Then, different algorithms allow to simplify the representation of a shadow and make measurements of parameters giving maximal geometrical information (as Fourier Descriptors, elongation, or compacity). Thus, we obtain for each shadow a vector of parameters directly usable by the discriminators. In a previous work [1], we investigated an efficient classification method. Thus, this task was performed with four classifiers in parallel (Bayesian classifier, K-nearest neighbours, Restricted Coulomb Energy method, **Multi-Layer Perceptron**). Results exhibited the best performances for **MLP** especially for poorly segmented images. In this study, we propose to reduce the complexity of the **MLP** while keeping a good generalization rate. Therefore, we develop a new pruning technique based on the Optimal Brain Damage (**OBD**) algorithm proposed by Le Cun et al. [2]

2. DATA BASES

The aim of the developed processes is the automatic classification of objects lying on the seafloor such as the cylinder on the image (a) of the table 1. To achieve this task we have to detect and extract the object cast shadow by per-

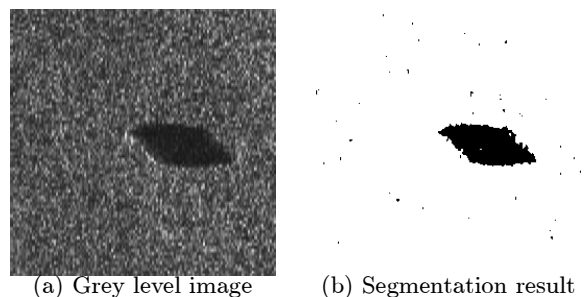(a) Grey level image   (b) Segmentation result

Table 1: Sonar image of a cylinder lying on the seafloor.

forming the segmentation of the image. The image (b) of the table 1 shows the segmentation result. This segmentation is good enough to allow the extraction of relevant geometrical parameters. Thus, each shadow is associated to a parameter vector. So the classifier is able to locate the sample in the representation space and gives it a label. With the collected images, we composed a training set and a test set. Four classes of man-made objects are represented corresponding to objects with the most common occurrence on the seafloor. The training set contains 976 images and the test set 600 images divided up as follows :

$$
\text{training set} \begin{cases} 326 & \text{rock} \\ 200 & \text{sphere} \\ 250 & \text{cylinder} \\ 200 & \text{cone} \end{cases}
$$

$$
\text{test set} \begin{cases} 200 & \text{rock} \\ 100 & \text{sphere} \\ 125 & \text{cylinder} \\ 175 & \text{cone} \end{cases}
$$

The **MLP** used in these tests contains 14 input neurons and 10 hidden neurons.

3. PRINCIPLE OF NEURAL NETWORK PRUNING

A basic problem in machine learning is to minimize the system complexity. This is important for two reasons: first, a low complexity system is faster and easier to implement on limited resources devices, and, second, it is now well-known that good generalization performances are associated with a minimal representation.

Various strategies have been proposed to simplify neural networks [5]. There are basically two groups of methods. With the first group, training and pruning are performed simultaneously: for example, an extra term that penalizes large weights can be added to the error function. When this term is the sum of the squares of the weights, the algorithm is known as weight decay. With the second group of methods, training is performed first, with an oversized neural network, and then pruning is done on the trained network by eliminating unnecessary weights. A measure of the importance of each weight has to be defined: it is generally based on a second order approximation of the error function. The algorithm used in this paper is based on this kind of strategy.

One of the most widely used algorithms is known as OBD (Optimal Brain Damage), and was proposed by Le Cun et al. [2]

Let us introduce the following notations:

$$\left\{ \begin{array}{ll} E_p & \text{the error for pattern } p \\ E_T = \sum_p E_p & \text{total error (training base)} \\ w_j & \text{the weight number } j \end{array} \right.$$

If the weights are changed by δw_j, a second order approximation of the variation of the error is given by:

$$\begin{array}{ll} \delta E_T & = \sum_j \frac{\partial E_T}{\partial w_j} \delta w_j + \frac{1}{2} \sum_j \frac{\partial^2 E_T}{\partial w_j^2} (\delta w_j)^2 \\ & + \frac{1}{2} \sum_{i \neq j} \frac{\partial E_T}{\partial w_i \partial w_j} \delta w_i \delta w_j \end{array}$$

When weight j is suppressed, the resulting variation of error is:

$$\delta E_T = -\frac{\partial E_T}{\partial w_j} w_j + \frac{1}{2} \frac{\partial^2 E_T}{\partial w_j^2} w_j^2 \qquad (1)$$

It is usually claimed that, since the network has been trained, is has reached a minimum, hence $\frac{\partial E_T}{\partial w_j} = 0$. Thus, the variation of error reduces to:

$$\delta E_T = \frac{1}{2} \frac{\partial^2 E_T}{\partial w_j^2} w_j^2 \qquad (2)$$

This value is called the saliency of the weight: it is a measure of the importance of the weight. It is estimated for each weight, and the weights with smallest saliencies are eliminated first. The method based on equation 2 is known as Optimal Brain Damage (OBD) [2] [3]. An improvement of OBD is Optimal Brain Surgeon (OBS) [4]: it automatically rescales the remaining weights, hence avoiding further training after the pruning phase.

4. NEW RESULTS ABOUT THE FIRST ORDER TERM

Influence of the first order on the estimated error

In [2][3], the first order term which appears in equation 1 is neglected: the justification is that the network has been trained, hence is has reached a minimum of the error and the gradient is null. However, we have observed that usually this is not true.

The method that we propose in the next section is based on the observation that the first order term in equation 1 is usually not negligible. Even if the network is supposed to have reached a minimum, it is never exactly on the minimum (this is due to the discrete nature of the back-propagation algorithm: there are always small oscillations around the minimum at the end of learning).

Figure 1 shows the true error (i.e. SSE obtained on the training set) and the estimated error with respect to the number of removed weights in the case where only the 2nd order is taken into account (using eq. 2). Notice that the saliency of each weight is computed after each removal and the weight with smallest saliency is removed. The network is not retrained during the pruning procedure. Figure 2 is the same as figure 1, but now the first order term is not neglected when the saliency is computed: equation 1 is used instead of equation 2. This figure shows that including the first order term in the saliency expression allows to improve the results i.e. the error curve lies in a lower level during almost all the pruning process. This improvement appears more clearly on the figure 3 where the true error curves obtained with equation 2 and equation 1 are drawn simultaneously. This result is confirmed with the figure 4 which shows the generalization rate with respect to the number of removed weights in the cases of the two previous equations.

However, figure 2 exhibits a quite large difference between the true and estimated errors. More precisely, in the case of equation 1 the error is under-estimated. Moreover, we notice that an unfortunate phenomenon occurs for the 92nd removed weight : the estimated error seems stable whereas the real error increases in a large range. This phenomenon becomes a real problem if we want to define a stop criterion. Indeed, a simple criterion would be expressed as follows : stop pruning when the estimated error has reached a minimum (usually, there is one). In the case of figure 2, the minimum would be obtain for the 95th removed weight but for this weight we can see on the figure 3 that the real error obtained with equation 2 would have been smaller.

After these observations, it appears that including the first order term should not be systematically required. To try to explain this result, which seems strange at a first glance (on an intuitive point of view, since there are less approximations in equation 1 than in equation 2, the estimation of the error should be better), we propose to draw the error with respect to modified values of a weight which has been selected with equation 1 to be removed. Two different examples will allow us to visualize the case where the first order term is relevant and the case where only the second order terms is interesting.

On the figure 5 we can see the real error curve and two parabolic approximation curves. The parameters of these approximation curves are obtained with the first and the second order derivatives which are computed at the point corresponding to the initial weight. If we observe the intersection points between these curves and the verti-
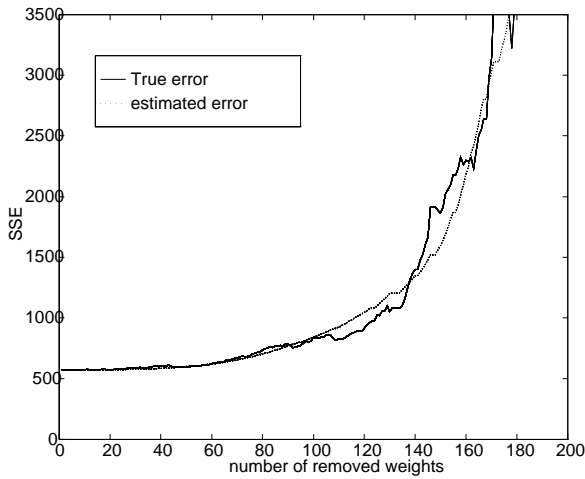
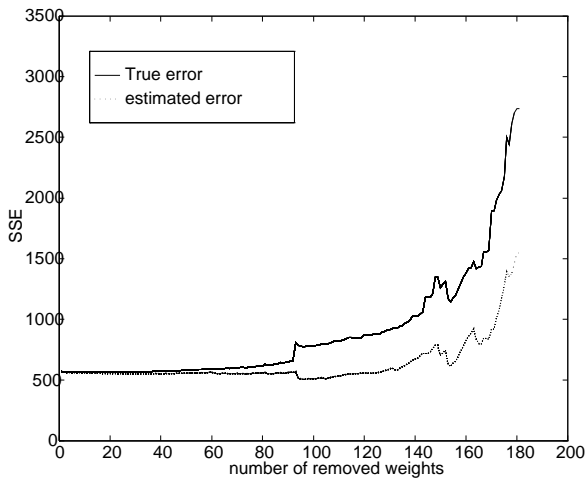Figure 1: True and estimated errors obtained with the 2nd order term.

Figure 2: True and estimated errors obtained with the 1st and 2nd order terms.
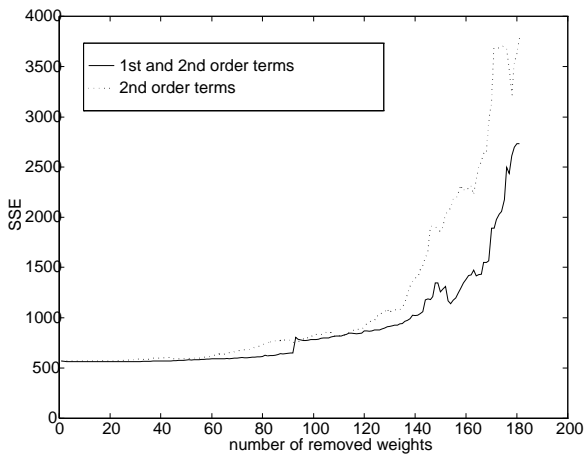

Figure 3: Real errors obtained with and without order 1.
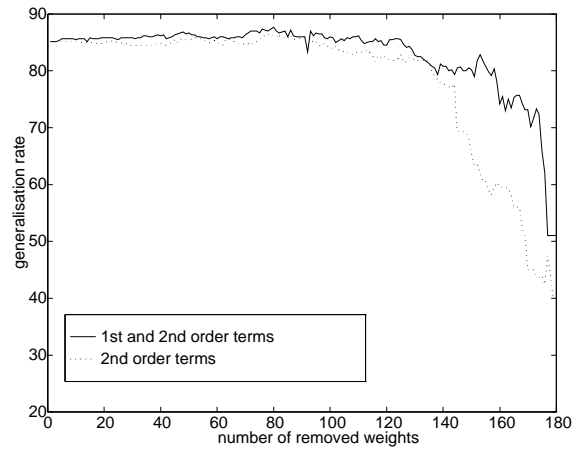

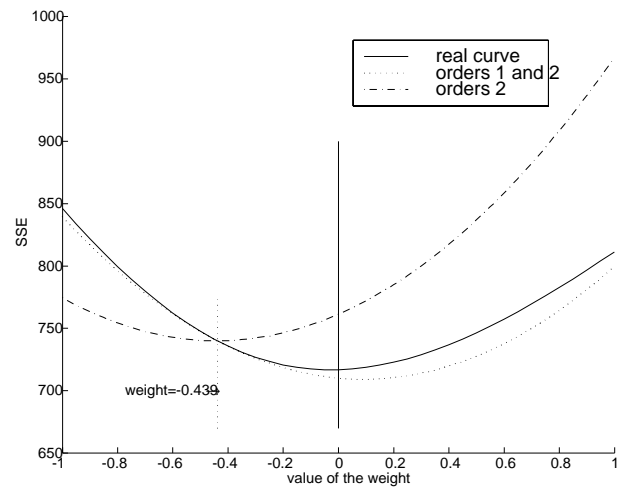Figure 4: Generalisation rates obtained with and without order 1.

Figure 5: True error and two parabolic approximations with respect to the value of the weight in the case where the first order is relevant. The value of the initial weight is $w_j = -0.439$.

cal axis, it is clear that neglecting the first order term can hardly be justified. Here, the first order estimation saliency allows to remove an unnecessary weight leading to a decrease of the error. On the other hand, on the figure 6 the saliency computed with equation 1 has a low value (the weight would be removed first) whereas the true saliency is very high. In this particular case, the saliency computed only with the second order term is better (i.e. closer of the real curve) because the local minimum of the error curve (m_j) is situated between zero and the initial weight (w_j). The left curvature of the real error yields an important increase of the error. This explains the divergence between the estimated error and the true error on the figure 2. In the next section we will propose a method that takes profit of the interest of the first order term in order to improve the performance of a pruned network.

Proposed pruning method

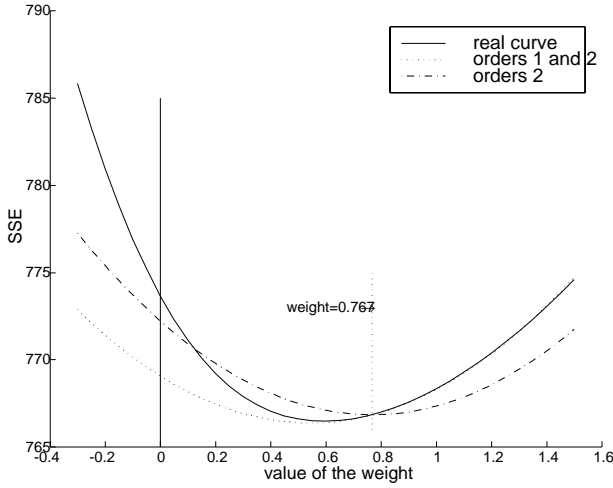The true justification of the removal of the first order

Figure 6: True error and two parabolic approximations with respect to the value of the weight in the case where the first order can be neglected. The value of the initial weight is $w_j = -0.767$.

term in [2] [3][4] is not the fact that it is negligible (in fact, usually it is not negligible at all), but the fact that it can degrade the estimation of the saliency for some weights. As we see in the previous section, using equation 1 leads to globally under-estimate the error. The analysis of figures such as figure 6 drawn at different steps during the pruning process shows that this phenomenon is maximum when the local minimum is between zero and the initial weight. In this configuration, it would be interesting to not systematically take into account the first order term. Therefore, the method we propose to determine the location of the initial weight with respect to the local minimum and, in the case where the local minimum is between zero and the initial weight, is based on the following observations:

- If the initial weight moves off the local minimum the value of the saliency given by equation 2 increases and becomes larger than the true saliency. In this case, it would be useful to include the first order term.

- If the initial weight moves forward the local minimum the employment of equation 2 avoids to under-estimate too much the error.

Now we have to determine the boundary value above which we take into account the first order term. Intuitively, we can consider that this value depends on the value of m_j. Indeed, if m_j decreases, we lead towards a situation where m_j is close to zero. In this case we have to take into account the first order term whatever the value of the initial weight. On the contrary, if the value of m_j increases the boundary has to increase too. Moreover, we choose to have a soft ponderation of the first order term in equation 1. So, the saliency can be express as follows:

$$\delta E_T = -P(w_j, m_j)\frac{\partial E_T}{\partial w_j}w_j + \frac{1}{2}\frac{\partial^2 E_T}{\partial w_j^2}w_j^2 \qquad (3)$$

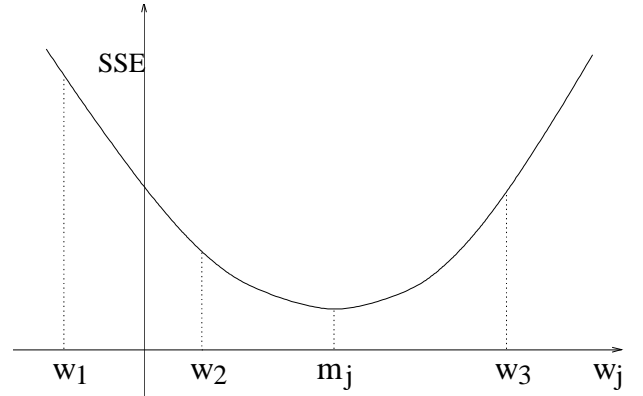$P(w_j, m_j)$ represents the ponderation applied to the first

Figure 7: Possible location of the initial weight with respect to the position of m_j.

order. To define it, let us consider the figure 7. It exhibits three possible types of locations for the initial weight with respect to m_j. If the weight is situated in w_1 we have seen that we wholly take into account the first order term. To detect this configuration we determine the sign of the product $w_j m_j$. If this sign is negative, w_j is in the same position as w_1.

If the weight is situated in w_2, we take into account the first order term too. Here, $w_j m_j$ is positive but $|w_j|$ is smaller than $|m_j|$.

As a matter of fact, the ponderation term should be 1 if $w_j m_j < 0$ or $|w_j| < |m_j|$.

Finally, for a weight situated such as w_3, the more w_j moves off m_j, the more we take into account the first order term. So, we use the following ponderation function :

$$P'(w_j, m_j) = 1 - e^{\frac{-(w_j - m_j)^2}{2\sigma_j^2}}$$

Where σ_j is determined in order to maintain a certain value of P' when $w_j = 2m_j$.

To sum up, the ponderation function becomes :

$$P(w_j, m_j) = \begin{cases} 1 & \text{if} \quad w_j m_j < 0 \\ 1 & \text{if} \quad |w_j| < |m_j| \\ 1 - e^{\frac{-(w_j - m_j)^2}{2\sigma_j^2}} & \text{elsewhere} \end{cases} \qquad (4)$$

The figure 8 shows that the equation 3 used with the relation 4 leads to improve the estimation efficiency. Especially, the problem of divergence encountered for the 92nd removed weight has disappeared. On the figure 9, we can notice that this new result is better than the one obtained with only the 2nd order, during all the pruning process. The generalization rate presents on the figure 10 leads to the same conclusion.

5. CONCLUSIONS

We have taken an active interest in the pruning method **OBD** developed by Le Cun. Especially, we have seeked to determine the real influence of the first order term in the estimation of the saliency. Tests performed in the scope of an object classification problem on sonar images have led
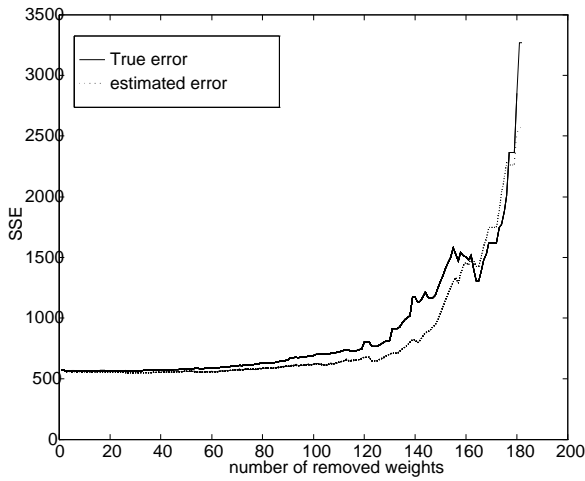
Figure 8: True and estimated errors obtained with the ponderation of first order term.
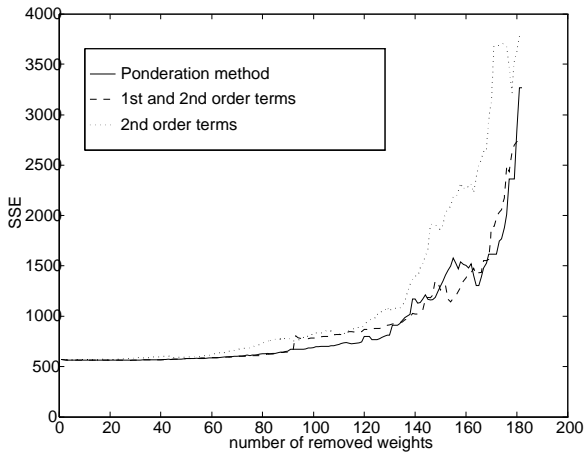

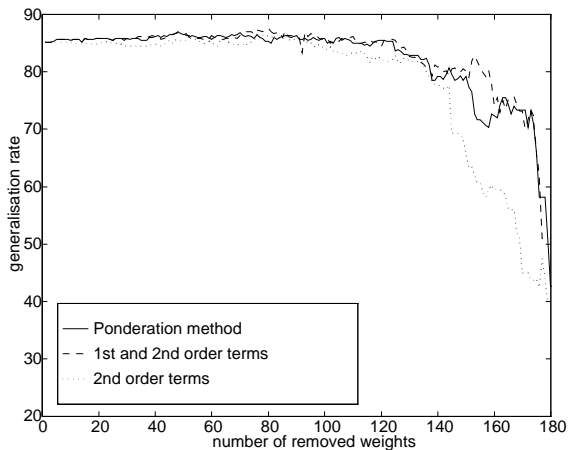Figure 9: Comparison of the true error curves for the three methods.

Figure 10: Comparison of the generalisation rate curves for the three methods.

to the conclusion that the first order term is generally not negligible but that its systematical use leads to divergence problems between the estimated and real errors. Using these observations, we have developed a pruning method based on the ponderation of the first order term which allows to clearly improve the generalization rate and to have a quite good estimation of the real error.

References

[1] P. Galerne, K. C. Yao, G. Burel, "Objects classification using Neural Network in sonar imagery", Proceedings of EUROPTO on New Image Processing Techniques and Applications, Spie Vol. 3101, pp. 306-314, 18-19 June 1997, Munich, FRG.

[2] Y. Le Cun, J. S. Denker, S. A. Solla, "Optimal Brain Damage", in Advances in Neural Information Processing Syst. II, ed. D. S. Touretzsky (Morgan Kaufman, San Mateo,1990), pp. 598-605

[3] J. Gorodkin, L. K. Hansen, A. Krogh, C. Svarer, O. Winther, "A quantitative study of pruning by Optimal Brain Damage", International Journal of Neural Systems, vol. 4, No. 2, June 1993, pp. 159-169.

[4] B. Hassibi, D. G. Stork, G. J. Wolff, "Optimal Brain Surgeon and General Network Pruning", in Proc, 1993 IEEE International Conference on Neural Networks, eds. E. H. Ruspini and al., vol. 1, pp. 293-299, San Francisco, 28 March - 1 April 1993.

[5] R. Reed, "Pruning Algorithms - A survey", IEEE Transactions on Neural Networks, vol. 4, No. 5, september 1993