Optimal Brain Surgeon Variants For Optimization

Mohammed Attik and **Laurent Bougrain** and **Frédéric Alexandre**[1]

Abstract. The determination of the optimal architecture of a multilayer perceptron (MLP) to solve a specific problem is an important and a difficult task. Several approaches based on saliency analysis, such as the Optimal Brain Surgeon method (OBS), have been developed in this field. Nevertheless, we show in this paper that OBS does not gives an optimal architecture. We also present the advantages of applying OBS on the architecture obtained by a variable selection method. New hybrid methods are proposed. A comparison of our approaches to standard techniques for architecture optimization is presented.Simulation results obtained on the Monk's problem illustrate the specificities of each method described in this paper.

1 Introduction

The objectives of optimization can be numerous: improving the prediction performance, providing faster predictor, providing a better understanding of the underlying process that generates data (providing variable selection, facilitating extraction of rules) and reducing the time and the cost to collect and transform data.

In the last years several heuristic methods based on computing the saliency for topology optimization have been proposed, e.g., Optimal Brain Damage (OBD) ([3]), or OBS ([2]). These methods are known as pruning methods. The main difference between these pruning methods for feature selection is the way of selecting the weight or the variable to eliminate. We are interested in this paper in variants based on OBS saliency calculation.

The first motivation of this work is to show that OBS gives a suboptimal architecture. We interested to maximise reduction of network complexity, i.e. removal of all the unecessary variables and weights.

The second motivation is to propose a new hybrid approach which combines variable selection method and OBS to improve the results of optimization.

2 Feature selection techniques

The OBS method was introduced by Hassibi and Stork ([2]) as a method to significantly reduce the number of weights in a neural network. Stahlberger and Riedmiller ([6]) proposed to the OBS's users, a calculation, called Generalized Optimal Brain Surgeon (G-OBS), to obtain in a single step the update to apply to every weight when deleting a subset of m weights. This calculation presents a combinatory calculation to know which weights should be deleted. An implementation for this method is proposed in ([1]) which defines the subset of connections by the smallest saliencies.

The OBS method has inspired some methods specialized in variable selection like Unit-Optimal Brain Surgeon (Unit-OBS)([6]), or recently Flexible-Optimal Brain Surgeon (F-OBS) and Generalized Flexible-Optimal Brain Surgeon (GF-OBS)([1]). The Unit-OBS was propsed by Stahlberger and Riedmiller ([6]), it is a variable selection algorithm, which computes, using the calculation G-OBS, which input unit will generate the smallest increase of error if it is removed. The F-OBS has been proposed in ([1]), its particularity is to remove connections only between the input layer and the hidden layer. The GF-OBS is a combination of F-OBS and G-OBS ([1]). Thus, this algorithm removes in one stage a subset of connections only between the input layer and the hidden layer.

A comparison between these different methods was studied for minimizing the network topology and for variable selection (see [6], [5], [4],[1] for more details).

3 Our methods for optimization

We propose two approaches for architecture optimization :

- The first approach uses a unique method specialized to optimization, OBS or G-OBS, applied several times consecutively until there is no weight to eliminate.
- The second approach is a hybrid approach, it is a methodology of eliminate variables/weights. In the first step this approach proposes to use a variable selection method and OBS in the second step. Stahlberger ([6]) suggests to combine Unit-OBS and OBS to obtain best results as compared to OBS. We suggest using variable selection methods, F-OBS and GF-OBS to optimize the architecture. The hybrid algorithm can be summarized by the following :

1. Choose a reasonable network architecture.
2. Choose a variable selection method (F-OBS, Unit-OBS, GF-OBS).
3. Execute the variable selection method.
4. Execute OBS or G-OBS (one or several times).

When we execute OBS, we use the new architecture obtained and we initialise all values of the weights again before training. In this work we interested to apply the OBS in the second step in the hybrid approach.

4 Experiments and results

In this section, we want to compare the different strategies of the OBS variants including the hybrid methods for minimizing the network topology. We use the first Monk's problem to evaluate the performance for each method. This well-known problem (See [7]) requires the learning agent to identify (true or false) friendly robots based on six nominal attributes.

[1] LORIA-INRIA, Campus Scientifique - BP 239, 54506 Vandœuvre-lès-Nancy, France, {Mohammed.Attik@loria.fr, Laurent.Bougrain@loria.fr, Frederic.Alexandre@loria.fr}

To forecast the class according to the 17 input values (one per nominal value coded as 1 or -1 if the characteristic is true or false), the MLP starts with 3 hidden neurons containing a hyperbolic tangent activation function. This number of hidden neurons allows a satisfactory representation able to solve this discrimination problem. The total number of weights for this fully connected network (including a bias) is 58. This value will be compared to the remaining weights after pruning.

After MLP training, the model is accepted if the mean of square error is ≤ 0.001 on both the training and the validation dataset, and if the performances in classification are equal to 100% according to the confusion matrix. This stopping criterion is also used by the pruning methods. In this study, GF-OBS and G-OBS remove three weights at the same time.

We select two values as measures of the performance for optimization: the number of preserved weights and the number of preserved variables. For each method, 500 different initializations were tested. Some results are presented as histograms in Table (2).

According to all obtained histograms, we notice that :

- The results of optimization are considerably improved when OBS or G-OBS are applied two times consecutively.
- The hybrid techniques give good results compared to the simple optimization methods OBS and G-OBS.
- There is a certain compromise between the number of variables and the number of weights for all methods. F-OBS appears best A to find this compromise.

Table (1) gives a classication of the different methods according to the best solutions or performances for two points of view : the minimal number of preserved weights and the minimal number of preserved variables.

Table 1. Summary of the model performances (1 is assigned to the best method, and so on.)

Methods	Minimal number of preserved weights	Minimal number of preserved variables
OBS method	6	6
G-OBS method	7	7
OBS method (twice)	1	4
G-OBS method (twice)	5	5
Unit-OBS hybrid method	4	1
F-OBS Hybrid method	2	2
GF-OBS Hybrid method	3	3

5 Conclusion

We have presented in this study different techniques based on the saliency calculation of OBS for architecture optimization. We used statistic methods to compare empirical performances of these different variants. We have shown that using OBS or G-OBS gives a sub-optimal architecture optimization. We have first proposed to apply OBS several times consecutively starting from the reduced architecture and reinitializing the weights until there is no more weight to eliminate. Then, we have presented a hybrid approach for optimization combining variable selection methods and OBS. Finally, in order to optimize artificial neural network architecture for rule selection, we recommend either to apply the F-OBS hybrid technique, because F-OBS algorithm better keeps the variables which are associated to rules to extract.

Acknowledgment

This work was supported in part by the BRGM (Bureau de recherches géologiques et minières).

REFERENCES

[1] M. Attik, L. Bougrain, and F. Alexandre, 'Optimal brain surgeon variants for feature selection', in *International Joint Conference on Neural Networks (IJCNN)*, (2004).

[2] Babak Hassibi and David G. Stork, 'Second order derivatives for network pruning: Optimal brain surgeon', in *Advances in Neural Information Processing Systems*, eds., Stephen José Hanson, Jack D. Cowan, and C. Lee Giles, volume 5, pp. 164–171. Morgan Kaufmann, San Mateo, CA, (1993).

[3] Yann Le Cun, John S. Denker, and Sara A. Solla, 'Optimal brain damage', in *Advances in Neural Information Processing Systems: Proceedings of the 1989 Conference*, ed., David S. Touretzky, pp. 598–605, San Mateo, CA, (1990). Morgan-Kaufmann.

[4] T. Ragg, S. Gutjahr, and H. Sa. Automatic determination of optimal network topologies based on information theory and evolution, 1997.

[5] Thomas Ragg, Heinrich Braun, and Heiko Landsberg, 'A comparative study of neural network optimization techniques', in *Proc. of the ICANNGA 97*. Springer-Verlag, (1997).

[6] Achim Stahlberger and Martin Riedmiller, 'Fast network pruning and feature extraction by using the unit-OBS algorithm', in *Advances in Neural Information Processing Systems*, eds., Michael C. Mozer, Michael I. Jordan, and Thomas Petsche, volume 9, p. 655. The MIT Press, (1997).

[7] S. B. Thrun, J. Bala, E. Bloedorn, I. Bratko, B. Cestnik, J. Cheng, K. De Jong, S. Džeroski, S. E. Fahlman, D. Fisher, R. Hamann, K. Kaufman, S. Keller, I. Kononenko, J. Kreuziger, R. S. Michalski, T. Mitchell, P. Pachowicz, Y. Reich, H. Vafaie, W. Van de Welde, W. Wenzel, J. Wnek, and J. Zhang, 'The MONK's problems: A performance comparison of different learning algorithms', Technical Report CS-91-197, Pittsburgh, PA, (1991).

Table 2. Distribution of the number of preserved variables (left side) and preserved weights (right side)