A Prior of Saliency Based Pruning Algorithms

Jon Sporring*

E-mail: sporring@diku.dk
Department of Computer Science, University of Copenhagen
Universitetsparken 1, DK-2100 Copenhagen
DENMARK
Phone: +45 35321400
Fax: +45 35321401

Technical Report: DIKU–97/8, ISSN 0107-8283
http://www.diku.dk

April 9, 1997

*Part of this work was performed during a Joint-Study visit at IBM, Almaden Research Center, 650 Harry Road, San Jose, CA 95120-6099, USA

Abstract

In saliency based pruning algorithms a pruning decision is made based on an ordering of the parameters. We will in this article focus on the fact that this ordering is invariant under certain transformations, and with that knowledge an equivalence class of algorithms is developed all yielding identical prunings. A sub-class is demonstrated to have a simple but sensible interpretation in terms of Bayesian decision theory.

1 Introduction

For some function classes, such as the feed forward neural networks, the number of parameters is very large when compared to the usual size of datasets to be fitted. To give an example, the number of parameters in the simplest universal feed forward network $f : \mathbb{R}^N \to \mathbb{R}^M$ (Hornik, 1989; Cybenko, 1989), the number of parameters grow as $d(M+N+1)$, where d is the number of internal nodes or hidden neurons as they are also called.

To reduce complexity and increase generalization, a function class can be analyzed by examining each individual parameter for it's importance. This process of removing parameters based on such an analysis is known as pruning and is the subject of this article. We will illustrate how a specific pruning scheme can be used to generate a similarity class of algorithms based on invariance. One popular algorithm, Optimal Brain Damage (OBD) (Cun et al., 1990), is interpreted in a statistical manner as a Maximum A Posteriori (MAP) or information theoretical code length functional, and it is thus shown how OBD can be interpreted in terms of the implicit prior on the feed forward network function class.

Before we begin, the reader should note that although the foundations of MAP and coding are very different, there is in the idealized code length setting applied here, a one to one correspondence between the two. Idealized code lengths are determined through Shannon's entropy-inequality (Rissanen, 1989) as,

$$L(\theta) \simeq -\log P(\theta), \tag{1.1}$$

where L is the code length for the particular entity θ and P is its corresponding probability. Under the assumption that P is known, there exists algorithms, such as the Huffman and especially the Arithmetic coding algorithm, that approach an equality of the above. Conversely, it is straight forward through the equality to design a probability distribution given a set of complete prefix codes. We are thus in this loose sense free to choose the formalism best suited for our needs.

2 Pruning

Fitting a function to a set of data points is often accomplished by minimizing an error function $E(\boldsymbol{\theta})$, where $\boldsymbol{\theta}$ is the set of parameters. The definition of saliency as we use it in this article is the increase in E when one or more parameters are removed, i.e. set to zero. The increase by removal of the parameter set $\{\theta_{i_1}, \ldots, \theta_{i_n}\}$ will be called $\Delta_{\{\theta_{i_1},\ldots,\theta_{i_n}\}}E$, and an ordering is thus induced,

$$\Delta_{\alpha_{i-1}}E \geq \Delta_{\alpha_i}E \geq \Delta_{\alpha_{i+1}}E \tag{2.1}$$

where we used the sloppy notation of α_j to denote a set of parameters. The exact pruning decision performed is not of importance to the work presented in this article, as long as the decision is based only on the ordering. Generally the set of parameter removals that generate the lowest increase in the error function is pruned.

The exact increase is often too computational expensive to evaluate, and for analytical error functions (and hence analytical functions) the ordering may be estimated by a

truncated Taylor series,

$$\begin{aligned}\Delta E(\boldsymbol{\theta}, \Delta\boldsymbol{\theta}) &\equiv E(\boldsymbol{\theta} - \Delta\boldsymbol{\theta}) - E(\boldsymbol{\theta}) \\ &= -\sum_i \frac{\partial E(\boldsymbol{\theta})}{\partial \theta_i}\Delta\theta_i + \frac{1}{2}\sum_i\sum_j \frac{\partial^2 E(\boldsymbol{\theta})}{\partial\theta_i\partial\theta_j}\Delta\theta_i\Delta\theta_j + \dots . \end{aligned} \tag{2.2}$$

To give an example, OBD uses the sum over all data points of the L_2 squared differences between the dataset and the function output and estimates the saliency to second order.

A mathematical as well as computational convenient restriction is to consider only single parameter prunings. This reduces the number of saliencies to be computed to equal the number of parameters (not yet pruned), and it simplifies the Taylor expansion to

$$\Delta_p E(\boldsymbol{\theta}, \Delta\boldsymbol{\theta}) = -\theta_p \frac{\partial E(\boldsymbol{\theta})}{\partial \theta_p} + \theta_p^2 \frac{1}{2}\frac{\partial^2 E(\boldsymbol{\theta})}{\partial \theta_p^2} + \dots , \tag{2.3}$$

for each parameter θ_p. Be reminded that in this case, $\Delta\boldsymbol{\theta} = [0, 0, \dots, 0, \theta_p, 0, \dots, 0]^T$.

Note that the key issue, when using a Taylor expansion, is that E and hence the function, must be analytical and the estimation point must be within the radius of convergence. The last point is generally ignored in the literature. Further, the ordering itself does not indicate when it is no longer reasonable to continue pruning. This must be determined by exterior constraints such as generalization maximization, see e.g. (Sporring, 1995; Svarer et al., 1993; Rasmussen, 1993) references therein and many others.

3 A Prior of Saliency Based Pruning Algorithms

For the simplicity of the following argument we will investigate single parameter pruning algorithms, but note that the results holds for multi parameter prunings as well. Assume that we have an ordering of the parameters such that,

$$\Delta_{p_{i-1}} E \geq \Delta_{p_i} E \geq \Delta_{p_{i+1}} E. \tag{3.1}$$

It is at once noticed that since monotonic transformation with positive slope preserve inequality, the above ordering is also unaffected,

$$T(\Delta_{p_{i-1}} E) \geq T(\Delta_{p_i} E) \geq T(\Delta_{p_{i+1}} E), \tag{3.2}$$

I.e. all continuous transformation $T : \mathbb{R} \to \mathbb{R}$ with $\partial_x T \geq 0$ for all x does not affect the pruning order. We will now study a linear transformation $T(\Delta E) = a\Delta E + \mathrm{b}$, for constants $a > 0$ and b, and show that the pruning algorithms described in this article can be interpreted in terms of a model expectancy.

Examine the following function,

$$L(\boldsymbol{\theta}) = \alpha E(\boldsymbol{\theta}) + \sum_i \beta_i \log|\theta_i| + \gamma, \tag{3.3}$$

where α, β_i, and γ are constants. α must be greater than or equal zero, and the set of β_i's must be chosen such that the saliency order is not disturbed. Generally we will assume that $\beta_i = 0$ when $\theta_i = 0$ and use the convention that $0 \log 0 = 0$.

Proposition 1. *For $\alpha > 0$ and a constrainted set of β_i's L preserves the pruning order of any analytical error function E in a Taylor series truncated to finite order.*

Proof. The proof is given in appendix A. □

Proposition 2. *For $\alpha > 0$ and a constrainted set of β_i's L is the only functional of any analytical error function E for which the chance of L is a linear function of the change of E in the Taylor series truncated to finite order.*

Proof. See appendix B for the proof. □

There are several key points to notice. First of all, the particular set of β_i's where $\beta_j = \beta$ for all j does not upset the pruning order. To see this, write the constraints on β_i as,

$$\alpha \frac{\Delta_{p_{i-1}} E - \Delta_{p_i} E}{\sum_{j=1}^{J} j^{-1}} - \beta_{i-1} \geq -\beta_i \geq \alpha \frac{\Delta_{p_{i+1}} E - \Delta_{p_i} E}{\sum_{j=1}^{J} j^{-1}} - \beta_{i+1} \tag{3.4}$$

where J is the truncation order. For identical β_j's the original order is retained.

Secondly, this particular choice of identical constants β_j's is precisely the limit for the truncation order going towards infinity, since the sum in the denominator will tend to infinity as J does, hence the band of different allowable β_j's will tend to zero, i.e. $\beta_j \to \beta$ for all j as $J \to \infty$.

Finally, if E is an analytical function then L is too. We have a semi-group property in the sense that we can define two sequential non-pruning disturbing extensions as in Equation 3.3 and get a third non-disturbing pruning. Thus define L' as,

$$L'(\boldsymbol{\theta}) = \alpha' L(\boldsymbol{\theta}) + \sum_i \beta_i' \log |\theta_i| + \gamma', \tag{3.5}$$

with a new set of constants chosen as prescribed previously, but this time based on L instead of E. This is of course just

$$L'(\boldsymbol{\theta}) = \alpha' \alpha E(\boldsymbol{\theta}) + \sum_i (\alpha' \beta_i + \beta_i') \log |\theta_i| + \alpha' \gamma + \gamma', \tag{3.6}$$

Again we see that the requirements to be fulfilled are

$$\alpha' \alpha \frac{\Delta_{p_{i-1}} E - \Delta_{p_i} E}{\sum_{j=1}^{J} j^{-1}} - \alpha' \beta_{i-1} - \beta_{i-1}' \geq -\alpha' \beta_i - \beta_i' \geq \alpha' \alpha \frac{\Delta_{p_{i+1}} E - \Delta_{p_i} E}{\sum_{j=1}^{J} j^{-1}} - \alpha' \beta_{i+1} - \beta_{i+1} \tag{3.7}$$

and for $\beta_j = \beta$ and $\beta_j' = \beta'$ this requirement is trivially fulfilled. Note that this is a different approach than choosing two different sets of β_i's both chosen from the same analytical function and then combined. This last approach is in general not a pruning order invariant.

We will now examine the choice of identical $\beta_j = 1$ for all j. Equation 3.3 can be interpreted in the coding setting as the sum of code lengths of the noise model and the parameter model, and in the MAP setting as minus the logarithm of the noise probability times the prior,

$$L = L(\mathcal{D}|\boldsymbol{\theta}) + L(\boldsymbol{\theta}) = -\log P(\mathcal{D}|\boldsymbol{\theta}) - \log P(\boldsymbol{\theta}), \tag{3.8}$$

where,

$$P(\mathcal{D}|\boldsymbol{\theta}) = \exp(-\alpha E(\boldsymbol{\theta}) - \gamma_0). \tag{3.9}$$

In the example of OBD E, is the sum over data points of the square of a $\mathcal{L}_2$ norm, and this can be interpreted as a normal product distribution with a unit standard deviation, and

$$P(\boldsymbol{\theta}) = \exp(\gamma_1) \prod_i |\theta_i|^{-\beta_i} \simeq \exp(\sum_i -\log\eta - \log\lfloor|\delta\theta_i|\rfloor - \log\log\lfloor|\delta\theta_i|\rfloor - \dots), \tag{3.10}$$

where $\gamma = \gamma_0 + \gamma_1$, η is a normalization constant, and δ is the discretization constant to truncating reals into integers. The sum is continued just until the repeated logarithm yields a negative number. This last equation is also known as Rissanen's Universal Distribution of Integers (Rissanen, 1989) and most clearly demonstrates the difference between coding and the MAP methods. While the MAP methodology is best suited for continuous distributions, such as Jeffrey's semi-prior $\gamma_1/|\theta_i|$ (Jaynes, 1968), the problems of normalization and discretization is much better handled in the coding methodology. The key difference between the two is that while Jeffrey's prior can only be implemented on a finite interval of the real axis in order for it to be normalized, Rissanen's distribution is normalizable for all countable sets like the set of all positive integers. Hence using Jeffrey's prior one is concerned with the interval size D in order to evaluate the normalization constant $\gamma_1 = \int_1^D 1/x\, dx$, while one's concern when using Rissanen's distribution is the discretization constant δ, i.e. the number of digits accounted for. It should be noted that there are of course other more sophisticated MAP and coding implementations of distributions for real numbers. These and other implementation issues concerning this coding prior can be found in (Sporring, 1995).

We may thus view the OBD pruning algorithm as a greedy algorithm searching for the minimum in Equation 3.8 by removing the least significant parameter through the error estimate. This will increase the actual error, but decrease the cost of the model.

4 Conclusion

This paper has demonstrated that a large class of saliency based pruning methods, where the saliency is calculated from analytical functions, can be used to generate a similarity class of pruning algorithms all having same pruning order. The (in a sense most) general extension in this similarity class is used to interpret OBD in terms of Bayesian Maximum A Posteriori (MAP) or code-length functionals and a Prior has thus been made explicit. This is found to be Jeffrey's Prior (Jaynes, 1968), which is a very natural un-committed result for the following reasons:

- Jeffrey's Prior is scale invariant in the sense that it assign equal probability mass to the intervals $1-10$, $10-100$, etc.. It is also the basis of what is known as Benford's law, which although surprising has been empirically validated on numerous datasets of surprisingly different nature e.g. (Buck et al., 1993).

- A very close relative, Rissanen's Universal Distribution of Integers is frequently used in the coding industries, and one can show (Rissanen, 1989) that it is an optimal code for large integers.

Finally we will conclude that although OBD uses poor estimates when the parameter values of the net are large, it is a good un-committed choice in the view of the scale invariant properties of the implicit prior.

5 Acknowledgments

I would especially like to thank Peter Johansen, Mads Nielsen, Luc Florack and Robert Maas for the many and enlightening discussions during this work.

A Proof of Proposition 1

We will now prove that the change of L (Equation 3.3) under certain restriction generates the same pruning order as the change of any analytical function E up to any but finite truncation order in the Taylor series.

The change of L can be written as,

$$\Delta L(\boldsymbol{\theta}, \Delta\boldsymbol{\theta}) \equiv L(\boldsymbol{\theta} - \Delta\boldsymbol{\theta}) - L(\boldsymbol{\theta}) = -\sum_i \frac{\partial L(\boldsymbol{\theta})}{\partial(\theta_i)}\Delta\theta_i + \frac{1}{2}\sum_{i,j}\frac{\partial^2 L(\boldsymbol{\theta})}{\partial(\theta_i)\partial(\theta_j)}\Delta\theta_i\Delta\theta_j + \ldots \tag{A.1}$$

Clearly, the mixed derivatives of the sum of the logarithms are zero, so we need only examine non-mixed terms. First we need to evaluate the n'th derivative of $\log|x|$. For simplicity write $\log|x|$ as $1/2\log x^2$, we will now prove by induction that

$$\frac{\partial^n}{\partial x^n}\frac{1}{2}\log x^2 = (-1)^{n-1}(n-1)!x^{-n}. \tag{A.2}$$

Assume that the n'th derivative is given as above. The $n+1$'th derivative is then

$$\frac{\partial}{\partial x}(-1)^{n-1}(n-1)!x^{-n} = (-1)^{n-1}(n-1)!(-n)x^{-n-1} = (-1)^n n!x^{-(n+1)}. \tag{A.3}$$

For $n = 1$, the first derivative is seen to be,

$$\frac{\partial}{\partial x}\frac{1}{2}\log x^2 = \frac{1}{x} = (-1)^0 0!x^{-1}, \tag{A.4}$$

thus completing the proof.

The j'th term in the Taylor expansion of L is given as,

$$(-1)^j\frac{\partial^j L(\boldsymbol{\theta})}{j!\,(\partial(\theta_p))^j}(\Delta\theta_p)^j \quad = \quad (-1)^j\alpha\frac{\partial^j E(\boldsymbol{\theta})}{j!\,(\partial(\theta_p))^j}(\Delta\theta_p)^j - \frac{\beta_p}{j\theta_p^j}(\Delta\theta_p)^j. \tag{A.5}$$

We identify the first term to be α times the identical term in the Taylor expansion of E, and further because of the symmetry, i.e. $\Delta\theta_p = \theta_p$, we quickly find that

$$\Delta_p L(\boldsymbol{\theta}, \Delta\boldsymbol{\theta}) = \alpha \Delta_p E(\boldsymbol{\theta}, \Delta\boldsymbol{\theta}) - \beta_p \sum_{j=1}^{J} j^{-1}, \tag{A.6}$$

up to any finite truncation order J. The β_j's are to be chosen such that the pruning order is maintained, i.e. since $\Delta_{p_{i-1}} E \geq \Delta_{p_i} E \geq \Delta_{p_{i+1}} E$ then so must $\Delta_j L$ and thus for positive α,

$$\alpha \frac{\Delta_{p_{i-1}} E - \Delta_{p_i} E}{\sum_{j=1}^{J} j^{-1}} - \beta_{i-1} \geq -\beta_i \geq \alpha \frac{\Delta_{p_{i+1}} E - \Delta_{p_i} E}{\sum_{j=1}^{J} j^{-1}} - \beta_{i+1}. \tag{A.7}$$

This completes the proof.

B Proof of Proposition 2

We will show that L of Equation 3.3 is the unique function that generates linear invariance to the change of any analytical function E.

A linear transformation of the change in error E must have the form,

$$\Delta L(\boldsymbol{\theta}, \Delta\boldsymbol{\theta}) = a \Delta E(\boldsymbol{\theta}, \Delta\boldsymbol{\theta}) + b, \tag{B.1}$$

where a and b are constants. We will now investigate the possible functions in the Taylor description for a and b.

The constant a is a scaling constant and it is trivially seen that if a is a function of $\boldsymbol{\theta}$ and $\boldsymbol{\Delta\theta}$ then the contribution can be eliminated by an opposite term in b. We will thus assume a to be a positive constant. The constant b is another matter. We are faced with the choice of a function h such that

$$L(\boldsymbol{\theta}) = aE(\boldsymbol{\theta}) + h(\boldsymbol{\theta}) + c \tag{B.2}$$

which in the Taylor series behaves such that

$$b = \sum_{j=1}^{J} (-1)^j \frac{\partial^j h(\boldsymbol{\theta})}{j! \, (\partial\theta_p)^j} (\theta_p)^j \tag{B.3}$$

is a constant for arbitrary but finite J. The first order terms constraint the problems to sums of functions of only one parameter. Thus either h is independent on θ_p or,

$$\frac{\partial^j h(\boldsymbol{\theta})}{(\partial\theta_p)^j} = b_p \frac{1}{(\theta_p)^j} \tag{B.4}$$

for any j and p, and constants b_p restricted as discussed in appendix A. Thus

$$h(\boldsymbol{\theta}) = \sum_i b_i \log|\theta_i| + b_0 \tag{B.5}$$

is the only solution for arbitrary constant b_0. This completes the proof.

References

Buck, B., Merchant, A. C., and Perez, S. M. (1993). An illustration of Benford's first digit law using alpha decay half lives. *European Journal of Physics*, 14:59–63.

Cun, Y. L., Denker, J., and Solla, S. (1990). Optimal brain damage. In Touretzky, D., editor, *Advances in Neural Information Processing Systems*, pages 598–605. San Mateo.

Cybenko, G. (1989). Approximation by superpositions of a sigmoidal function. *Mathematics of Control, Signals and Systems*, 2:303–314.

Hornik, K. (1989). Multilayer feedforward networks are universal approximators. *Neural Networks*, 2:359–366.

Jaynes, E. T. (1968). Prior probabilities. *IEEE Transactions on systems science and cybernetics*, 4(3):227–241.

Rasmussen, C. E. (1993). Generalization in neural networks. Master's thesis, Technical University of Denmark.

Rissanen, J. (1989). *Stochastic Complexity in Statistical Inquiry*. World Scientific.

Sporring, J. (1995). Pruning with minimum description length. In *Proceedings of the 5. Scandinavian Conference on Artificial Intelligence (SCAI'95)*, pages 157–168, Trondheim, Norway.

Svarer, C., Hansen, L. K., and Larsen, J. (1993). On design and evaluation of tapped-delay neural network architectures. In *et al.*, H. B., editor, *Proceedings of the 1993 IEEE Int. Conference on Neural Networks (ICNN93)*, pages 45–51.