ESANN'2007 proceedings - European Symposium on Artificial Neural Networks
Bruges (Belgium), 25-27 April 2007, d-side publi., ISBN 2-930307-07-2.

Complexity Bounds of Radial Basis Functions and Multi-Objective Learning

Illya Kokshenev and Antônio P. Braga

Universidade Federal de Minas Gerais - Depto. Engenharia Eletrônica
Av. Antônio Carlos, 6.627 - Campus UFMG Pampulha 30.161-970,
Belo Horizonte, MG - Brasil

Abstract. In the paper, the problem of multi-objective (MOBJ) learning is discussed. The problem of obtaining apparent (effective) complexity measure, which is one of the objectives, is considered. For the specific case of RBFN, the bounds on the smoothness-based complexity measure are proposed. As shown in the experimental part, the bounds can be used for Pareto set approximation.

1 Introduction

Similarly to other universal approximators, neural networks are capable to yield good generalization by minimizing a fitting criterion when the amount of observations is sufficiently large. However, non-synthetic data is most likely to be finite and disturbed by noise, leading single objective error minimization learning approaches often to poor generalization and overfitting. This kind of behaviour led to the development of generalization control methods, which aim at obtaining a proper balance between bias and variance, by also minimizing network complexity [1]. Pruning [2] and regularization [3] methods are commonly used to achieve that goal. While pruning algorithms control complexity by manipulating network structure, regularization aims at controlling network output response with penalty functions. However, error and network complexity (structural or apparent) in these two approaches are treated as single objective problems. Although this may result on good generalization models, they are highly dependent on user defined training parameters. In addition to that, it is well known that error and complexity are conflicting objectives and, similarly to bias and variance, demand balancing instead of joint minimization. This viewpoint of learning demands multi-objective (MOBJ) methods [4], which treat empirical and structural risks as two independent objectives.

The MOBJ problem can be defined by introducing error and complexity objective functions $\phi_e(\omega)$ and $\phi_c(\omega)$, repectively. With the first one representing the empirical risk and the second one representing the structural risk, we may now formulate MOBJ learning as the vector optimization problem

$$\min_{\omega\in\Omega}(\phi_e(\omega), \phi_c(\omega)), \tag{1}$$

where ω is the vector of network parameters in the parameter space Ω.

Since the objectives are conflicting in the region of interest, the solution of (1) is a Pareto-optimal front $\Omega^* \subseteq \Omega$, in which the elements $\omega^* \in \Omega^*$ satisfy the conditions $\forall\omega : \{\phi_e(\omega) \geq \phi_e(\omega^*), \phi_c(\omega) \geq \phi_c(\omega^*)\}$. In other words, the optimization

problem results in the optimal solutions that represent the best compromise between the two objectives. It means that for every solution $\omega \notin \Omega^*$, there are others in Ω^* that have lower complexity and error.

Usually, the squared error criterion and the norm of network weights $\|w\|$ are taken as the error and complexity measures for MLP networks [5]. A general definition for assessing apparent complexity of other network types is not known, what is an obstacle for MOBJ learning. In this paper we present the smoothness-based apparent complexity measure and propose its bounds for RBF network. We show that the apparent complexity is limited by $\sigma^{-1}\|w\|_1$, where σ is the radius of the hidden layer functions and $\|\cdot\|_1$ is the Manhattan norm operator ($1-$norm). It is shown also that this result can be used to control generalization of RBF with MOBJ learning.

2 Apparent complexity measure for RBFN

Considering neural network as the function f in Sobolev spaces $\mathbb{W}^{k,p}$, its smoothness could be represented in terms of the Sobolev [6] norm

$$\|f\|_{k,p} = \sum_{i=0}^{k} \|f^{(i)}\|_p = \sum_{i=0}^{k} \Big(\int |f^{(i)}(t)|^p \, dt \Big)^{1/p}. \tag{2}$$

Since only the second or higher order elements of sum (2) represent nonlinear properties, the element $\|f^{(2)}\|_2$ can be chosen for nonlinear smoothness criterion, which is also considered as a penalty function in regularization [3, 7]. The apparent complexity measure based on $\|f\|_{2,2}$ may be therefore expressed as

$$\phi_c(\omega) = \int \Big\| f''(x,\omega) \Big\| \, \partial x = \int \|\triangle_x f\| \, \partial x, \tag{3}$$

where $f(x,\omega)$ is the mapping function of the neural network with parameters ω.

Let's consider the particular case of n-input single output RBF network containing m Gaussian functions of the same radius σ and prototype matrix $c = (c_1, c_2, \ldots, c_m)$. Introducing the centered input vector $\delta_j = \frac{x - c_j}{\sigma}$ with respect to j-th prototype and the centered kernel function $K(\delta) = exp(-\frac{1}{2}\delta^2)$, the network output can be written as

$$f(x, w, \sigma, c) = \sum_{j=1}^{m} w_j K(\delta_j), \tag{4}$$

where w is the $(m \times 1)$ output weights vector.

Hence, the radial basis functions with spherical receptive fields is considered, the Hessian of (4) is diagonal and its Euclidean norm in (3) is

$$\|\triangle_x f\| = \left(\sum_{i=1}^{n} \left(\sum_{j=1}^{m} \sigma^{-2} w_j K\left(\delta_j\right) \left(\delta_j^2 - 1\right) \right)^2 \right)^{\frac{1}{2}}.$$

According to the triangle property of the norm operator, we obtain the following inequality:

$$\|\triangle_x f\| \leq \sum_{j=1}^{m} \left(\sum_{i=1}^{n} \left(\sigma^{-2} w_j K\left(\delta_j\right) \left(\delta_j^2 - 1\right) \right)^2 \right)^{\frac{1}{2}}$$

or

$$\|\triangle_x f\| \leq \sigma^{-2} \sum_{j=1}^{m} |w_j| K\left(\delta_j\right) \left(\|\delta_j\|_4^4 - 2\|\delta_j\|^2 + n \right)^{\frac{1}{2}} \tag{5}$$

after simplification. One can see, that there are positive functions on both sides of (5). Thus, the inequality remains the same after integration, providing

$$\phi_c(w, \sigma, c) \leq \sigma^{-2} \sum_{j=1}^{m} |w_j| \int K\left(\delta_j\right) \left(\|\delta_j\|_4^4 - 2\|\delta_j\|^2 + n \right)^{\frac{1}{2}} \partial x.$$

Introducing $\Psi(\delta_j) = K\left(\delta_j\right) \left(\|\delta_j\|_4^4 - 2\|\delta_j\|^2 + n \right)^{\frac{1}{2}}$ and passing to the integral by $\partial\delta_i = \sigma^{-1}\partial x$, we obtain

$$\phi_c(w, \sigma, c) \leq \sigma^{-1} \sum_{j=1}^{m} |w_j| \int \Psi(\delta_j) \partial\delta_j.$$

Since $\Psi(\delta_j)$ does not depend on the network parameters, we may treat $\int \Psi(\delta_j)\partial\delta_j$ just as function of n and take it out from the sum. Accordingly, we obtain the bound on apparent complexity

$$\phi_c(w, \sigma, c) \leq \sigma^{-1} \|w\|_1 \Theta(n), \tag{6}$$

where $\Theta(n) = \int \Psi(\delta)\partial\delta$ and $\|\cdot\|_1$ is the Manhattan norm operator (1−norm).

3 MOBJ learning

Usually, the vector optimization problem (1) does not have an analytical solution, however in practice it is enough to approximate the Pareto front Ω^* with a finite number of solutions. According to MOBJ learning concepts, after approximation is obtained, the resulting "best" solution must be selected from Ω^* on a decision making step, in accordance to a *posteriori* criterion, such as validation error, maximum entropy *etc.*

When nothing is known about convexity of the Pareto front, the ϵ-constraint method can be used to achieve the approximation of Ω^*. The apparent complexity bound (6) is obtained, regardless of the prototype location. Here, we consider the case when prototypes are once determined by an appropriate strategy and does not participate in parameter search. Thus, inequality (6) leads to the ϵ-restrict approximation

$$\begin{cases} [w_i^*, \sigma_i^*] = \arg\min \phi_e(w, \sigma, c), \\ \sigma^{-1}\|w\|_1 \leq \epsilon_i. \end{cases} \tag{7}$$

In view of that, the constraints imposed on $\phi_c(\omega)$ bounds are more important than its magnitude, $\Theta(n)$ is neglected under assumption on its convergence.

4 Experiments

The experiment was concerned on the regression problem of *sinc* function

$$y(x) = \frac{sin(\pi x)}{\pi x} + \gamma, \tag{8}$$

where γ is the normally distributed zero mean and constant variance noise component. Since the problem is simple by itself, we treat it under conditions of high noise level and small number of observations in order to force the difficulty of obtaining good generalization.

We choose $m = 30$ radial basis functions and noise variance of $\sigma^2_{noise} = 0.2^2$. The training and validation sets were generated respectively by selecting 100 and 50 samples of (8) on the interval $x \in (0, 4\pi]$ normalized to $[0, 1]$. The test sample of (8) was taken without noise. Since the problem is one-dimensional, the prototype centers were equidistantly spaced on the input range.

The Pareto front was approximated for complexity bound restriction magnitudes $\epsilon_i \in [0, 80]$. The candidate solution for the corresponding ϵ_i was selected with respect to a minimal training error value calculated for the radius $\sigma \in [0.1, 0.5]$. According to (7), the w corresponding to a solution under chosen restrictions must satisfy $\|w\|_1 \leq \sigma\epsilon_c$. Therefore, for given ϵ_i and σ the output layer weight vector w was obtained by solving the constrained least squares problem

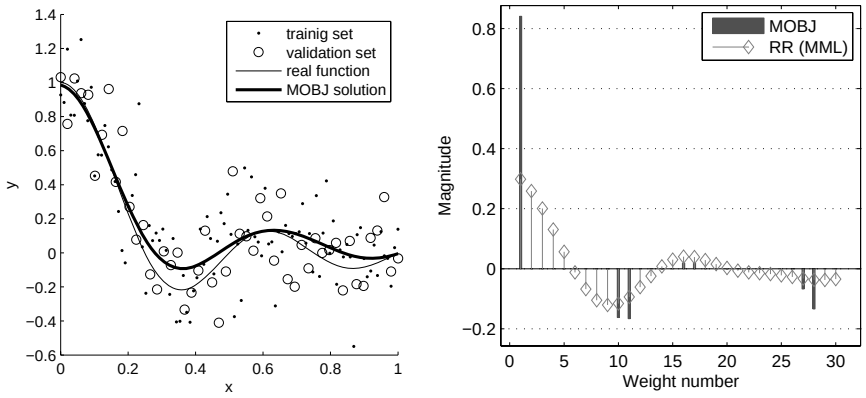
$$\begin{cases} E = \dfrac{1}{2}(Y - Hw)^2, \\ \|w\|_1 \leq \sigma\epsilon_i, \end{cases} \tag{9}$$

over the training set of N samples $(x(k)|y_d(k))$ using ellipsoid method. Here H is the $(N \times n)$ matrix of radial basis function values $h_{kj} = K(\delta_j(x(k)))$ and $Y = (y_d(1), y_d(2), \ldots, y_d(N))^T$ stands for the desired output vector.

The final solution, minimum of validation error, is picked-up within the obtained candidates set. The results of the Pareto front approximation are shown on Figure 2, where each auxiliary curve is also a Pareto front of the subproblem (9). The regression results are presented in Figure 1 and Table 1. Also, the comparison with the ridge regression (RR) method for various model selection criteria is given: Bayesian information criterion (BIC), generalized cross-validation (GCV) and maximum margin likelihood (MML).

5 Conclusions

The experimental results confirmed the efficiency of generalization control based on the proposed bounds. The solutions obtained by application of RR learning is close to the MOBJ results. In contrast to the RR results, most of the weights have exactly zero magnitude (see Figure 1 (b)), so the network structure can be



a) b)

Fig. 1: The regression results for MOBJ solution (a) and the weight magnitudes (b) in a comparison with RR results.

Solution	σ	$\|w\|_1$	$\sigma^{-1}\|w\|_1$	MSE (train/valid./test)
RR (GCV)	0.22	3.84	17.6	0.0405 / 0.0343 / 0.0048
RR (BIC)	0.22	3.97	17.6	0.0407 / 0.0349 / 0.0054
RR (MML)	0.16	1.95	11.9	0.0406 / 0.0329 / 0.0044
MOBJ	0.14	1.41	9.7	0.0403 / 0.0339 / 0.0042

Table 1: The experimental results.

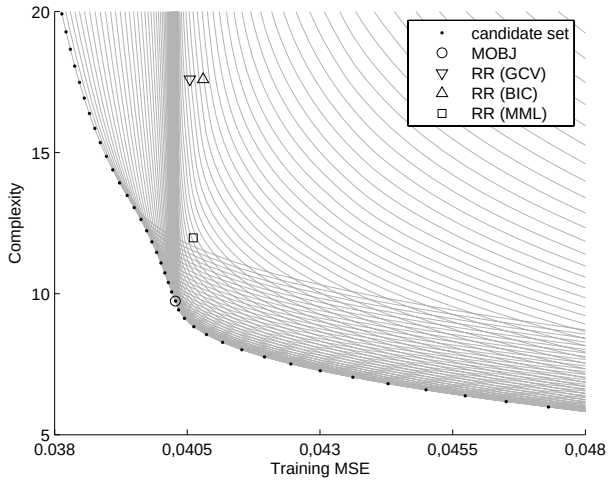


Fig. 2: The results of Pareto front approximation and the candidate set.

simplified without any loss. Hence, we can conclude that the proposed MOBJ approach involves both the pruning and regularization properties.

From the experimental results, it was observed that, under certain conditions,

the Pareto front presents non-convex intervals. Moreover, the best solution may belong to them.

It is noteworthy, that learning with regularization in a general nonlinear form is equivalent to solving the multi-objective problem (1) by weighting method, that is the minimization of the convex combination of the objectives $\min \phi_e(\omega) + \lambda\phi_c(\omega)$, where $\lambda \geq 0$ is the regularization parameter. For various values of λ

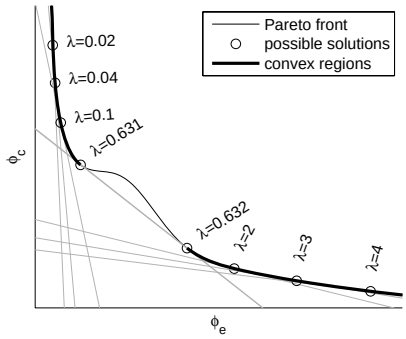


Fig. 3: The example of regularization in case of non-convex Pareto front.

the solutions will always form a convex front, which will concur with the Pareto front only on its convex intervals as it is shown on example Figure 3. Hence, we infer that regularization learning based on smoothness penalty (3) does not reach the best possible solutions from the non-convex Pareto front regions, so MOBJ learning is needed in such situations.

The proposed bounds on apparent complexity (6) for RBFs provide a new possibility for MOBJ learning. The concept can also evolve to more general RBF networks and other architectures, what brings also interest for further research.

References

[1] S. Geman, E. Bienenstock, and R. Doursat. Neural Network and the Bias/Variance Dilemma. *Neural Computation*, 4:1–58, 1992.

[2] R. Reed. Pruning algorithms - a survey. *IEEE Transactions on Neural Networks*, 4(5):740–746, 1993.

[3] Federico Girosi, Michael Jones, and Tomaso Poggio. Regularization theory and neural networks architectures. *Neural Computation*, 7(2):219–269, 1995.

[4] A. P. Braga, R. H. C. Takahashi, M. A. Costa, and R. A. Teixeira. Multi-objective algorithms for neural-networks learning. In Y. Jin, editor, *Multi-Objective Machine Learning (Series: Studies in Computational Intelligence)*, volume 16, pages 151–171. Heidelberg: Springer Verlag, 2006.

[5] R. A. Teixeira, A. P. Braga, and R. R. Saldanha R. H. C. Takahashi. Improving generalization of MLPs with multi-objetive optimization. *Neurocomputing*, 35:189–194, 2000.

[6] R. A. Adams and J. J.F Fournier. *Sobolev spaces*. Academic press, New York, second edition, 2003.

[7] M. A. Kon and L. Plaskota. Complexity of Regularization RBF Networks. In *Proceedings of International Joint Congress on Neural Networks, INNS*, pages 342–346, Washington, 2001.