

# Selection of important input variables for RBF network using partial derivatives

Jarkko Tikka and Jaakko Hollmén

Helsinki University of Technology  
Department of Information and Computer Science  
P.O. Box 5400, FI-02015 HUT, Finland  
tikka@mail.cis.hut.fi, <http://www.cis.hut.fi/tikka/>

**Abstract.** In regression problems, making accurate predictions is often the primary goal. Also, relevance of inputs in the prediction of an output would be valuable information in many cases. A sequential input selection algorithm for Radial basis function (SISAL-RBF) networks is presented to analyze importances of the inputs. The ranking of inputs is based on values, which are evaluated from the partial derivatives of the network. The proposed method is applied to benchmark data sets. It yields accurate prediction models, which are parsimonious in terms of the input variables.

## 1 Introduction

The goal of a regression problem is to learn an input-output relationship from data. Dependencies between the inputs and the output are typically nonlinear, and the exact functional form is unknown. Neural networks are widely utilized in regression problems, since they are relatively fast to train [11] and capable to approximate a wide class of functions accurately [5]. The disadvantage of neural networks is, that they include all the input variables and importances of inputs are unclear. We propose a backward input selection algorithm for RBF networks. The inputs are dropped one at a time from the model based on the ranking calculated from the partial derivatives. The resulting subsets of inputs are assessed using leave-one-out (LOO) error. The rejection of unimportant inputs increases the interpretability of the network, it may improve the generalization capability, and it also decreases the computational complexity of the final network [2]. The proposed algorithm can be seen as a wrapper input selection method [3]. Another approaches are filter [13] and embedded methods [12].

## 2 Radial basis function networks

Let us assume that we have  $N$  measurements from an output  $y_j$  and  $d$  inputs  $\mathbf{x}_j = [x_{j1}, \dots, x_{jd}]$ ,  $j = 1, \dots, N$ . The output of RBF network with a Gaussian basis functions is

$$\hat{y}_j = \sum_{m=1}^M \alpha_m K(\mathbf{c}_m, \mathbf{x}_j) + \alpha_0, \text{ where } K(\mathbf{c}_m, \mathbf{x}_j) = \exp\left(-\frac{\|\mathbf{c}_m - \mathbf{x}_j\|^2}{\sigma_m^2}\right), \quad (1)$$

and  $M$ ,  $\mathbf{c}_m$ , and  $\sigma_m$  are the number, the centers, and the widths of the basis functions [4], respectively. The model can also be written in the matrix form as

$\hat{\mathbf{y}} = \mathbf{K}\boldsymbol{\alpha}$ , where the elements of matrix  $\mathbf{K}$  are defined as  $\mathbf{K}_{jm} = K(\mathbf{c}_m, \mathbf{x}_j)$  and the  $(M + 1)^{\text{th}}$  column is the vector of ones corresponding to the bias term  $\alpha_0$ .

We place the Gaussian basis function on each training data point  $\mathbf{x}_j$  and set the widths of the basis functions to an equal value  $\sigma_m = \sigma$ . The parameters  $\boldsymbol{\alpha}$  are estimated by minimizing the regularized mean squared error (MSE)

$$J = \frac{1}{N} \sum_{j=1}^N (y_j - \hat{y}_j)^2 + \gamma \sum_{m=0}^N \alpha_m^2, \quad (2)$$

where the second term controls smoothness of the nonlinear mapping.

The generalization capability of the model is measured using the LOO error. For the fixed value of the width  $\sigma$ , the LOO error is the function of the regularization parameter  $\gamma$

$$\text{MSE}_{\text{LOO}}(\gamma) = \frac{1}{N} \mathbf{y}^T \mathbf{P} (\text{diag}(\mathbf{P}))^{-2} \mathbf{P} \mathbf{y}, \quad (3)$$

where  $\mathbf{P} = \mathbf{I}_N - \mathbf{K}(\mathbf{K}^T \mathbf{K} + \gamma \mathbf{I}_M)^{-1} \mathbf{K}^T$  and  $\text{diag}(\mathbf{P})$  is of the same size and has the same diagonal as  $\mathbf{P}$  but is zero off the diagonal [4]. In the optimization of  $\gamma$ , we use the golden section line search method [1]. The parameters  $\boldsymbol{\alpha}$  are found by minimizing Eq. (2) using the parameters  $(\sigma, \gamma)$ , which minimize Eq. (3).

### 3 Input variable selection algorithm

We propose a relevance measure to rank importance of each input variable in the model in Eq. (1). Relevance of the inputs  $x_i$ ,  $i = 1, \dots, d$ , can be measured using the partial derivatives of the output  $\hat{y}$  with respect to  $x_i$  [10, 9, 8]. The derivatives of the most relevant inputs vary most through the range of input values. The partial derivative of the RBF network with respect to  $x_i$  is

$$d_{ji} = \frac{\partial \hat{y}_j}{\partial x_{ji}} = \frac{2}{\sigma^2} \sum_{m=1}^M \alpha_m K(\mathbf{c}_m, \mathbf{x}_j) (c_{mi} - x_{ji}), \quad i = 1, \dots, d, \quad j = 1, \dots, N. \quad (4)$$

We use an Add10 data set as an example. It includes inputs  $x_i$ ,  $i = 1, \dots, 10$ , which are sampled independently from an uniform distribution  $\mathcal{U}(0, 1)$ . The output is  $y = 10 \sin(2x_1 x_2) + 20(x_3 - 0.5)^2 + 10x_4 + 5x_5 + \varepsilon$ , where  $\varepsilon$  is the Gaussian noise with zero mean and unit variance. All the variables were scaled to have zero mean and unit variance before training of a RBF network using  $N = 250$  samples. On the first column of Fig. 1, the partial derivatives of the RBF network with respect to the inputs  $x_3$  and  $x_4$  are shown. The values  $d_{j4}$  are almost constant, thus the dependency between  $x_4$  and  $\hat{y}$  is linear. The dependency between  $\hat{y}$  and  $x_3$  is quadratic, since  $x_3$  and  $d_{j3}$  are linearly dependent.

The median of the values  $d_{j3}$  is nearly zero, because of the cancellations between negative and positive values. Thus, the absolute values  $|d_{j3}|$  and  $|d_{j4}|$  might be more representative, which are shown as a function of  $x_{j3}$  and  $x_{j4}$  on the second column of Fig. 1. The histograms of  $|d_{j3}|$  and  $|d_{j4}|$  are presented on

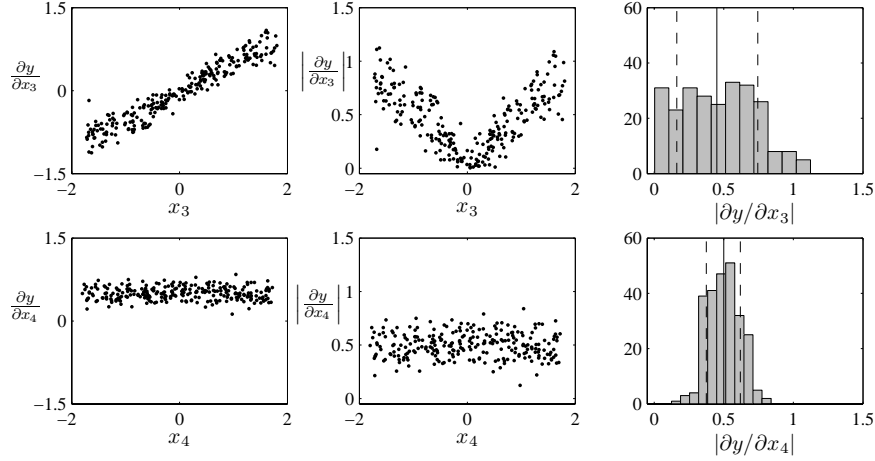


Fig. 1: Scatter plots of partial derivatives  $d_{j3}$  and  $d_{j4}$  (left) and absolute values  $|d_{j3}|$  and  $|d_{j4}|$  (middle) as a function of  $x_{j3}$  and  $x_{j4}$ , the histograms of the absolute values  $|d_{j3}|$  and  $|d_{j4}|$  (right), the solid lines are the medians and between the dashed lines is the central 67% quantile (right).

the third column. However, the medians of  $|d_{j3}|$  and  $|d_{j4}|$  are nearly equal and the relevance of the variables cannot be distinguished based on them. Thus, we propose to define a relevance measure of the input  $x_i$  as follows

$$r_i = m_{d_i} + \Delta_{d_i} , \quad (5)$$

where  $m_{d_i}$  is the median of the absolute values  $|d_{ji}|$ ,  $j = 1, \dots, N$ . The second term  $\Delta_{d_i}$  measures the variability of the values  $|d_{ji}|$ . It is defined as a difference  $\Delta_{d_i} = |d_{ji}|^{high} - |d_{ji}|^{low}$ , where  $|d_{ji}|^{high}$  and  $|d_{ji}|^{low}$  are the  $0.835N^{\text{th}}$  and  $0.165N^{\text{th}}$  values in the ordered list of the  $N$  absolute values  $|d_{ji}|$ . With previous choices, the difference  $\Delta_{d_i}$  is twice as large as the standard deviation in the case of the normal distribution. The larger the variation  $\Delta_{d_i}$  the more sensitive the output is to the corresponding input. The relevant inputs should also have clearly nonzero medians  $m_{d_i}$ . Thus, the most relevant input has the highest value for the relevance measure  $r_i$ . Both the median  $m_{d_i}$  and the difference  $\Delta_{d_i}$  are insensitive to the outliers in the data. Here, we use equal weighting for the two terms in (5), but unequal weighting could be used as well. Other relevance measures based on the partial derivatives are presented in [10, 14, 6].

We propose a backward input selection algorithm based on the relevance measure  $r_i$ . The algorithm starts by evaluating the RBF network with all the available input variables  $x_i$ ,  $i = 1, \dots, d$ . The hyperparameters  $\sigma^2$  and  $\gamma$  are selected by minimizing the LOO error. After that, the parameters  $\alpha$  are obtained as a solution to a system of linear equations in minimization of Eq. (2). The next step is to delete the least relevant input  $x_i$ , which is the input having the smallest value for the relevance measure  $r_i$ . The previous steps are repeated using the

---

**Algorithm 1** SISAL-RBF

---

- 1: Let  $\mathcal{L}$  be the set of the inputs  $x_i$ ,  $i = 1, \dots, d$
  - 2: Minimize the LOO error using the inputs in  $\mathcal{L}$ 
    - width of the basis functions  $\sigma$  fixed
    - optimize regularization parameter  $\gamma$  by minimizing (3)
  - 3: Repeat step 2. with various values of  $\sigma$ . Select the pair  $(\sigma_{\mathcal{L}}, \gamma_{\mathcal{L}})$ , which minimize Eq. (3)
  - 4: Use the pair  $(\sigma_{\mathcal{L}}, \gamma_{\mathcal{L}})$ , minimize the function in Eq. (2) with respect to  $\alpha$
  - 5: Evaluate relative importances of the inputs  $r_i$ ,  $i \in \mathcal{L}$
  - 6: Delete the input  $x_i$ , which has the smallest value for the relevance measure  $r_i$ , from the set of inputs  $\mathcal{L}$
  - 7: If  $\mathcal{L} \neq \emptyset$  go to step 2, otherwise go to step 8
  - 8: Select the set of inputs  $\mathcal{L}_v$ , which gives the smallest value for the LOO error
- 

Name	Training ( $N_t$ )	Test ( $N_{test}$ )	inputs	range of $\sigma^2$
Add10 <sup>1</sup>	250	9542	10	[1, 500]
Bank <sup>1</sup>	500	7692	32	[1, 10 <sup>4</sup> ]
Boston housing <sup>1</sup>	400	106	13	[1, 500]
Wine <sup>2</sup>	94	30	256	[5, 10 <sup>6</sup> ]

Table 1: Properties of the data sets.

remaining inputs, which results to the evaluation of  $d$  subsets of inputs. The final set of inputs minimize the LOO error. The sequential input selection algorithm for the RBF network (SISAL-RBF) is summarized in detail in Algorithm 1.

## 4 Experiments

SISAL-RBF was applied to four benchmark data sets (Add10, Bank, Boston housing, and Wine). In the case of Add10 data, the assessment of input selection results is straightforward, since the correct inputs are known. The data sets were randomly divided to the training and test sets. The sample sizes and the number of inputs are reported in Table 1. LOO errors were evaluated using 50 values of  $\sigma^2$ , which were equally spaced on a logarithmic scale in the ranges shown in Table 1. All the inputs and the outputs were scaled to have zero mean and unit variance to make the relevance measures comparable.

A forward selection (FS) algorithm was used as a baseline method to compare the performance of the proposed input selection strategy, since it is known that FS is robust against overfitting [7]. In the case of  $d$  inputs,  $(d + 1)d/2$  subsets of inputs have to be evaluated. FS could be stopped before all the inputs are

---

<sup>1</sup>Available from: <http://www.cs.toronto.edu/~delve/data/datasets.html><sup>2</sup>Available from: <http://www.dice.ucl.ac.be/mlg/index.php?page=DataBases>



