

MULTI-OBJECTIVE TRAINING OF RBF NETWORKS FOR LARGE DATA SETS WITH LMI'S

ELIZABETH F. WANNER[*], GLADSTON J. P. MOREIRA[†], EDUARDO G. CARRANO[‡], RICARDO H. C. TAKAHASHI[§], LUIZ H. DUCZMAL[¶]

[*]*Departamento de Matemática*
Universidade Federal de Ouro Preto
Ouro Preto, MG, Brasil

[†]*Departamento de Ciências Exatas*
Universidade Federal do Vale do Jequitinhonha e Mucuri
Teófilo Otoni, MG, Brasil

[‡]*Centro Federal de Educação Tecnológica de Minas Gerais*
Belo Horizonte, MG, Brasil

[§]*Departamento de Matemática*
Universidade Federal de Minas Gerais
Belo Horizonte, MG, Brasil

[¶]*Departamento de Estatística*
Universidade Federal de Minas Gerais
Belo Horizonte, MG, Brasil

Emails: efwanner@iceb.ufop.br, gladston@ufvjm.edu.br, egcarrano@deii.cefetmg.br, taka@mat.ufmg.br, duczmal@est.ufmg.br

Abstract— This work presents a Linear Matrix Inequality based training procedure for RBF networks that allows dealing with very large data sets. The proposed solution avoids a matrix inversion that is necessary in traditional approaches, thus increasing the dimension of the problems that can be dealt. The multi-objective setting for dealing with the bias-variance problem can be directly incorporated within the proposed methodology.

Keywords— Radial Basis Function, Linear Matrix Inequality, Multi-objective Optimization.

Resumo— Este trabalho apresenta uma metodologia de treinamento para uma rede RBF baseada em Desigualdades Lineares Matriciais que permite lidar com conjuntos extensos de dados. A metodologia proposta evita uma operação de uma inversão de matriz, necessária na metodologia tradicional, aumentando assim a dimensão dos problemas que podem ser resolvidos com a nova metodologia. O tratamento multiobjetivo para lidar com o problema de bias e variância pode ser diretamente incorporado à nova metodologia.

Keywords— Funções de Base Radiais, Desigualdade Linear Matricial, Otimização Multiobjetivo.

1 Introduction

Artificial Neural Networks (ANN) represent a technology that is applied to many fields such as non-linear phenomena modeling, time series analysis, signal processing, pattern recognition, etc. An important property is on the basis of such applicability: the ability to learn and generalize from input data. This makes it possible for ANN to solve complex problems that are difficult to treat via other methodologies.

Most of the problems that have been dealt by ANN's, up to now, involve the learning over a set of training data that is of moderate size: such sets have hardly presented more than one thousand input-output pairs. However, as the ANN's find new applications in fields such as on-line process control, on-line failure detection, and others, the size of training sets can grow easily, and the unavailability of training algorithms that are able to deal with large data sets becomes an important limitation that constrains the ultimate performance that is reachable by an ANN.

The RBF networks are particularly suitable for being trained with large data sets, since the training residues are linear in relation to the parameters to be adjusted (provided that the basis functions have already been chosen). The training procedure, in this case, becomes the resolution of a standard linear least-squares. The numerical procedure for solving this problem involves the inversion of a matrix that has dimension $N \times N$, where N denotes the length of the data vector. Once this inversion has been performed, the network weight vector (that is the result of the training procedure) is found via a matrix product. This procedure for RBF training can deal with data sets that are substantially larger than in the case, for instance, of the multilayer perceptron networks (MLP's), since the training of MLP's involves a non-linear optimization procedure. However, the training of RBF's reaches its limit of data set size when the matrix inversion procedure reaches the computer memory limits. This note presents a reformulation of the least-squares procedure in terms of a convex optimization with Linear Matrix Inequality (LMI) constraints. This re-formulation states the same

problem in a framework that involves a smaller memory requirement, allowing larger data sets. Once in the LMI format, the problem can be solved by any of the LMI solvers that are available. In this note, the SeDuMi is employed. SeDuMi is one of the most efficient LMI solvers that are available, being based on an interior-point self-dual convex cone optimization machinery and, as noticed in (Sturm, 99), it allows problems with relatively large sizes to be solved. SeDuMi is also publicly available, see *http://sedumi.mcmaster.ca/*.

It should be noticed that the proposed re-formulation directly allows the multi-objective training, as proposed in (Teixeira et al., 2000). The multi-objective training is performed in order to deal with noisy data, avoiding both data underfitting and data overfitting. The multi-objective training algorithms perform the balance between bias and variance by tradding-off the sum of the squared training error and the norm of the weight vector, see (Teixeira et al., 2000). This multi-objective training is directly performed within the LMI formulation, by simply adding a constraint in the optimization problem.

2 Radial Basis Function Network and the Multi-objective Approach

Radial basis functions are simple functions which decrease (or increase) monotonically with the distance from a central point, and that can be scaled and moved in order to form families of linearly independent functions. These functions can be used in order to build a basis for spaces of functions (for instance, the $\mathcal{C}_\infty$ space). A finite number of basis functions defines a finite-dimensional subspace of the function space. The trained RBF network can be considered as a projection of a function onto such finite-dimensional subspace.

A typical radial basis function that is used in RBF networks is the Gaussian function:

$$h(x) = exp\left(-\frac{(x-c)'(x-c)}{r^2}\right). \qquad (1)$$

The function parameters are the center c and the radius r.

In principle, radial functions could be employed in any sort of model (linear and nonlinear) and any sort of network (single-layer or multi-layer). However, since Broomhead and Lowe's seminal paper (Broomhead and Lowe, 1988), RBF networks have traditionally been associated with radial functions in a single-layer network. A RBF network is nonlinear if the basis function can move or change size or if there is more than one hidden layer. Here we foccus on single-layer RBF networks with functions which are fixed in position and size.

In this way, if the linear model is

$$f(x) = \sum_{j=1}^{m} w_j h_j(x) \qquad (2)$$

and the training set is $\{(\mathbf{x}_i, \hat{y}_i)\}_{i=1}^{p}$, $f(x)_i \in \mathbb{R}$, for all i, we need to minimize the sum of squared errors

$$S = \sum_{i=1}^{p} (\hat{y}_i - f(\mathbf{x}_i))^2 \qquad (3)$$

with respect to the weights of the model. The minimization of the norm of the vector $\mathbf{w} = [w_1, \ldots, w_m]'$ is also necessary, in order to avoid the model overfitting.

In a matrix form, the multi-objective problem can be described as

$$\mathbf{w}^* = \min_{\mathbf{w}} \begin{cases} f_1 = (\hat{Y} - H\mathbf{w})'(\hat{Y} - H\mathbf{w}) \\ \\ f_2 = \|\mathbf{w}\| \end{cases} \qquad (4)$$

where $\hat{Y} \in \mathbb{R}_{p\times 1}$ is a column matrix containg $(\hat{y}_i)_{i=1}^{p}$, $H \in \mathbb{R}_{p\times m}$ is the design matrix given by

$$H = \begin{bmatrix} h_1(\mathbf{x}_1) & h_2(\mathbf{x}_1) & \ldots & h_m(\mathbf{x}_1) \\ h_1(\mathbf{x}_2) & h_2(\mathbf{x}_2) & \ldots & h_m(\mathbf{x}_2) \\ \vdots & \vdots & \ddots & \vdots \\ h_1(\mathbf{x}_p) & h_2(\mathbf{x}_p) & \ldots & h_m(\mathbf{x}_p) \end{bmatrix} \qquad (5)$$

and $\mathbf{w} \in \mathbb{R}_{m\times 1}$ is the vector of network weights.

The first step when using multi-objective optimization is to obtain the Pareto-optimal set (Ehrgott, 2000), which contains the set of efficient solutions of the multi-objective problem. The next step selects the most appropriate solution within the Pareto-optimal set. There are several methods for handling multi-objective optimization problems (Ehrgott, 2000). In this work, the ϵ-constraint method was adopted (Haimes et al., 1971), since it can be directly accomodated within the LMI formulation.

Using the ϵ-constraint method, the multi-objective problem is redefined as a mono-objective problem of minimization of one of the objectives with a constraint ϵ in the other objective:

$$\begin{aligned} &\mathbf{w}^* = \min_{\mathbf{w}} (\hat{Y} - HW)'(\hat{Y} - HW) \\ &\text{s.t: } \|\mathbf{w}\| \leq \epsilon \end{aligned} \qquad (6)$$

Different Pareto-optimal solutions can be found by using different values of the constraint ϵ. In the next section we present the re-formulation of (6) as the optimization of a linear function with LMI constraints.

3 Linear Matrix Inequality

Linear Matrix Inequalities (LMIs) (Boyd et al., 1997) have emerged as powerful tools in areas such as control engineering, systems identification and structural design. A wide variety of problem can be formulated using LMIs and, once stated in such terms, the problem can be solved *exactly* by efficient convex optimization algorithms (LMIs solvers) based on interior-point algorithms. These solvers are significantly faster than classical non-linear optimization algorithms, and can deal with problems with larger sizes.

We will employ the following lemma to derive our methodology:

Lemma 1 (Schur's Lemma) *The statements (7) and (8) bellow are equivalent*

$$\left[\begin{array}{cc} Q & S \\ S' & R \end{array} \right] > 0 \tag{7}$$

$$\left\{ \begin{array}{l} R > 0 \\ Q - SR^{-1}S' > 0 \end{array} \right. \tag{8}$$

in which R and Q are symmetric matrices, S is a matrix with compatible dimensions, and $(\cdot) > 0$ denotes that the matrix argument is positive definite.

Proof: The proof of Schur's Lemma can be found in (Boyd et al., 1997). □

A direct result, based on the Schur's Lemma, is stated as:

Lemma 2 *Consider the following optimization problem with a quadratic objective function and a quadratic constraint:*

$$\begin{array}{l} \vec{x}^* = \arg\, \min\, (\vec{x} - \vec{x}_0)' Q (\vec{x} - \vec{x}_0) \\ \textit{s.t.} \ \left\{ \ (\vec{x} - \vec{x}_1)' H (\vec{x} - \vec{x}_1) - 1 \leq 0 \right. \end{array} \tag{9}$$

The optimization problem (9) can be re-stated as:

$$\vec{x}^* = \arg_{\vec{x}} \min_{\vec{x},\epsilon} \ \epsilon$$

$$\textit{s.t.} \left\{ \begin{array}{l} \left[\begin{array}{cc} \epsilon & (\vec{x} - \vec{x}_0)' \\ \vec{x} - \vec{x}_0 & Q^{-1} \end{array} \right] > 0 \\ \\ \left[\begin{array}{cc} C_1 & (\vec{x} - \vec{x}_1)' \\ \vec{x} - \vec{x}_1 & H^{-1} \end{array} \right] > 0 \end{array} \right. \tag{10}$$

Proof: Replace:

$$\min\, (\vec{x} - \vec{x}_0)' Q (\vec{x} - \vec{x}_0)$$

by:

$$\begin{array}{l} \min\ \epsilon \\ \text{s.t. } (\vec{x} - \vec{x}_0)' Q (\vec{x} - \vec{x}_0) < \epsilon \end{array}$$

The remainder operations are direct applications of Schur's Lemma to the quadratic inequality. □

Using Schur's Lemma, the mono-objective problem (6) can be re-written as

$$\mathbf{w}^* = \arg_{\mathbf{w}} \min_{\mathbf{w},\gamma} \ \gamma$$

$$\text{s.t.} \left\{ \begin{array}{l} \left[\begin{array}{cc} \gamma & (\hat{Y} - H\mathbf{w})' \\ (\hat{Y} - H\mathbf{w}) & I_p \end{array} \right] > 0 \\ \\ \left[\begin{array}{cc} \epsilon & \mathbf{w}' \\ \mathbf{w} & I_m \end{array} \right] > 0 \end{array} \right. \tag{11}$$

where I_p and I_m are identity matrix in $\mathbb{R}_{p \times p}$ and $\mathbb{R}_{m \times m}$ respectively.

The optimization problem (11) can be efficiently solved with any LMI solver based on interior point methods. In this work, we used SeDuMi (Sturm, 99) to solve this problem.

It is worthwhile to notice that the application of the weighted sum method (Ehrgott, 2000) to the problem (4) leads to a general Tikhonov regularization functional $L(w) = f_1 + \lambda f_2$. It is a convex combination of the objective functions and λ is the regularization parameter. For linear models, the unimodality of $L(w)$ is assured and all the Pareto optimal solution can be generated using proper values for λ. This property holds in this case if the basis functions are considered to be fixed, since the functional becomes convex in relation to the weight variable w.

4 Examples

4.1 Function Approximation

In this example, the function

$$y(x) = (1 + x - 2x^2) \exp(-x^2) + \gamma \tag{12}$$

in which γ is normally distributed zero mean and 0.2 standard deviation was approximated. A 30 RBFs was used and the training and validation sets were generated, respectively, by selecting 100 and 50 samples of y in the interval $[-4, 4]$. The centroids of the radial basis layer were determined from the training set using the well-known k-means clustering algorithm (MacQueen, 1967). The final solution was the one with a minimal error in the validation set. Figure 1 shows the chosen solution regarding this additional criterion. Figure 2 shows the Pareto set obtained via the multi-objective approach for solving the proposed RBF-LMI. The red dot represents the selected RBF which was the one with a minimal error in the validation set. Figure 3 shows all the obtained Pareto front approximations. It is possible to say that the proposed methodology presents a good generalization without overfitting.

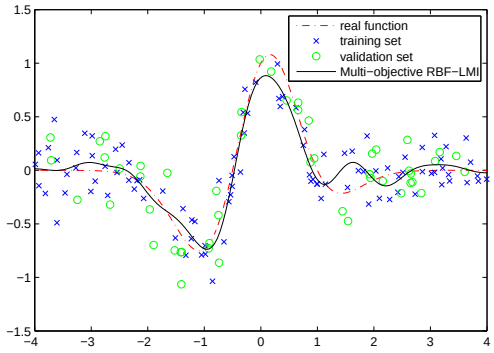

Figure 1: Function approximation given by the chosen RBF-LMI. The training and validation sets are also shown.

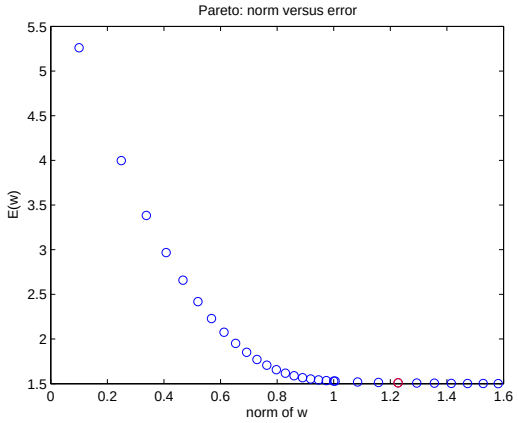

Figure 2: Pareto set obtained using the multi-objective approach for solving the RBF-LMI.

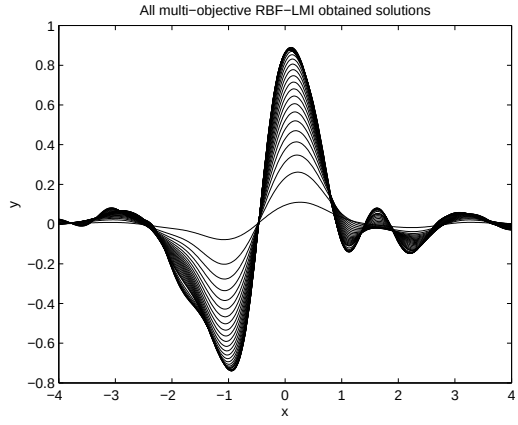

Figure 3: All the Pareto optimal solutions obtained via the multi-objective RBF-LMI approach.

4.2 Delineating Spatial Clusters

A problem of 2-D function approximation is presented now. Multiple or irregularly shaped spatial clusters are often found in disease or syndromic surveillance maps. We show a method to delineate the contours of spatial clusters, especially when there is not a clearly dominating primary cluster, through the methodology presented above. The method may be applied either for maps divided into regions or point data set maps. The spatial scan statistic (Kulldorff, 1997) is the usual measure of the strength of a cluster.

The methodology presented, using RBF and LMI, is trained as a function that approximates the spatial scan statistic for the whole map domain. The solution delineates regions of the map belonging to distinct level curves values, and those constitute the estimates of the primary and the several possible secondary clusters. For maps divided into m regions, the scan statistic is evaluated for each region taken individually, where each region is identified by the geographic coordinates of its centroid. We start defining a RBF + LIM artificial neural network with training set size m. The geographic coordinates and the scan values are, respectively, the net input and the desired output. A RBF + LMI, with 50 radial basis functions, is trained. Following the training phase, the scan function evaluation is extended for the whole domain of geographic coordinates. The areas lying inside the level curves above a certain threshold value are thus considered the most likely clusters. The balance between bias (rigidity) and variance (flexibility) of the net is obtained by means of its sizing (Teixeira et al., 2000). Larger topologies structures allow more flexibility but less bias. A similar approach is used for point data set maps.

The proposed algorithm deals with multi-objective optimization. The approximation for the function is constructed through minimization of a bi-objective problem: mean squared error and norm of weight vector. Using this methodology, the algorithm looks, in the RBF solution space, for balanced solution between bias and variance, that means between overfitting and underfitting.

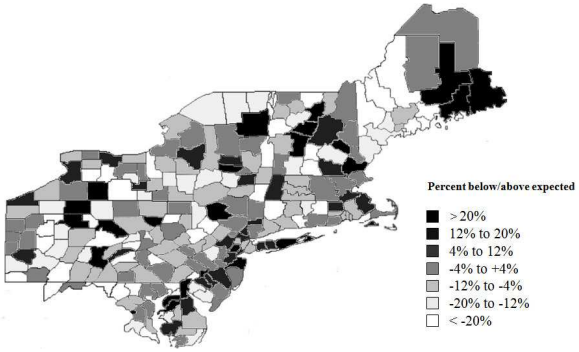

Figure 4: Northeast US counties map with incidence of female breast cancer in 1995 (age adjusted).

For obtaining the Pareto set of the problem (4), we use 20 values of ϵ in the formulation (6), where $\epsilon \in [0.01 \;\; 3.5]$. We draw the scan function level curves for breast cancer incidence in the Northeast United States in 1988-1992 (Kulldorff et al., 1997), see figure 4. Figures 5 and 6 show the scan function level curves, for $\epsilon = 0.01$ and $\epsilon = 3.5$ respectively. We obtain smoother clusters in Figure 5, when less variance and more bias are allowed. In Figure 6 we obtain otherwise sharply delineated, more irregularly shaped clusters, indicating less bias, due to the larger variance allowed.

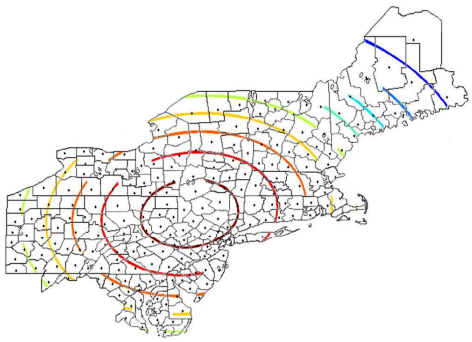

Figure 5: Level curve for Pareto point that corresponding to $\epsilon = 0.01$.

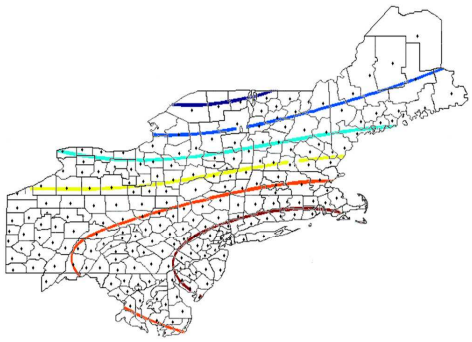

Figure 6: Level curve for Pareto point that corresponding to $\epsilon = 3.5$. For this value, this is the network which gives the small error on the validation set.

5 Conclusions

A new learning algorithm for improving generalization of RBF networks, which is based on multi-objective optimization, was presented. The multi-objective problem, which aims to minimize both the sum of squared error and the norm of weight vectors, is transformed into a mono-objective problem using ϵ-constrait method. The mono-objective problem is a quadratic one and can be solve via linear matrix inequliaty formulation.

We show that the RBF method presented is a fast and flexible algorithm for spatial cluster delineation. The irregularity of the cluster shapes can be adjusted varying the parameters of the neural network.

Acknowledgment

The authors acknowledge the support of the Brazilian agencies CAPES, CNPq and FAPEMIG.

References

Boyd, S., El-Ghaoui, L., Feron, E. and Balakrishnan, V. (1997). *Linear Matrix Inequalities in System and Control Theory*, SIAM.

Broomhead, D. S. and Lowe, D. (1988). Multivariate functional interpolation and adaptative networks, *Complex Systems* **2**: 321–355.

Ehrgott, M. (2000). *Multicriteria Optimization*, Spring-Verlag.

Haimes, Y. Y., Lasdon, L. and Wismer, D. (1971). On a bicriterion formulation of the problems of integrated system identification and system optimization, *IEEE Trans. on System, Man and Cybernetics* **1**(3): 296–297.

Kulldorff, M. (1997). A spatial scan statistic, *Comm. Statist. Theory Meth.* **26**: 1481–1496.

Kulldorff, M., Feuer, E., Miller, B. and Freedman, L. (1997). Breast cancer clusters in the northeast united states: a geographic analysis, *American Journal of Epidemiology* **146**: 161–170.

MacQueen, J. B. (1967). Some methods for classification and analysis of multivariate observations„ *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, University of California Press, pp. 281–297.

Sturm, J. F. (99). Using SeDuMi 1.02, a MATLAB toolbox for optimization over symmetric cones, *Optimization Methods and Software* **11-12**: 625–653.

Teixeira, R. A., Braga, A. P., Takahashi, R. H. C. and Saldanha, R. R. (2000). Improving generalization of MLPs with multi-objective optimization, *Neurocomputing* **35**(4): 189–194.