HYBRID TRAINING OF RADIAL BASIS FUNCTION NETWORKS IN A PARTITIONING CONTEXT OF CLASSIFICATION

Latifa OUKHELLOU, Patrice AKNIN

French National Institute for Transport and Safety Research.
INRETS-LTN, 2 av. Malleret-Joinville. 94114 Arcueil Cedex, France
Tel : 33 1 47 40 73 37. Email : aknin@inrets.fr

Abstract: The design of radial basis function networks (RBF) is rather complex because of the great number of parameters that must be adjusted: positioning and number of kernels, choice of the distance type and centre widths, weight values. This article details these points in the framework of classification tasks with a partitioning approach: the global K-class problem is split into K 2-class sub-problems. An adaptation of the Orthogonal Least Square method is presented in order to select the centres of each sub-classifier in connection with a particular stopping criterion based on the addition of a random centre. Moreover, different choices of distance and centre widths are compared and illustrated by a 4-class problem in the Non Destructive Evaluation domain.

Keywords: Radial Basis Function, Selection, Classification, Orthogonal Least Square, Mahalanobis Distance

Acknowledgement: this research is supported by the French Research Ministry within the framework of PREDIT Program (Research, Development and Innovation in Transport Systems).

1. Introduction

Neural networks are parsimonious approximators and they can achieve classification tasks by estimating the posterior probabilities of class membership [1]. In multilayer perceptron networks, made up of scalar product neurons, the discrimination frontiers between classes combine several hyperplans of the input space. In the RBF networks, made up of distance neurons, the classes are delimited by junctions of several ellipsoids. In this way, the RBF networks give more localized responses which show interesting performances, especially in supervised mode with outspread training data base [2].

After the introduction of the formalism of Gaussian RBF, a partitioning approach of the global K-class problem is presented. Then, a hybrid training of the RBF networks is tackled; the choice of the position and number of centres is presented in the linear regression context, associated with different stopping criteria. Four couples (distance type - centre width) are treated and compared in terms of classification performances.

All these particular points are illustrated by a Non Destructive Evaluation application which consists of the classification of eddy current signatures of metro rail defects into four classes.

2. Formalism

The output S of a single Gaussian RBF neuron is defined by the following vectorized equation:

\[ S = \Psi \left( \| \mathbf{X} - \mathbf{C} \|_A \right) = \exp \left( - \frac{\| \mathbf{X} - \mathbf{C} \|^2}{2\beta^2} \right) \] (1)

\( \mathbf{X} \) is the input vector, \( \mathbf{C} \) the centre of the neuron, \( \Psi \) the activation function and \( \beta \) the centre width.

Figure 1 : model of a single distance neuron

The distance operator is considered in its generalized form:

\[ \left( \mathbf{X} - \mathbf{C} \right) \]
\[ \|X - C_i\|^2 = (X' A X)' \]  \hspace{0.5cm} (2)

where A is a positive defined normalization matrix.

So, the complete expression of the Gaussian RBF neuron output is,

\[ S = \exp\left(-\frac{1}{2\beta^2}(X - C)' A (X - C)\right) \]  \hspace{0.5cm} (3)

We can recognize the formulation of a multidimensional Gaussian distribution with centre \( C \) and \( \beta^2A^{-1} \) as variance-covariance matrix.

3. RBF network structure for classification

RBF networks are made up of 2 layers: the hidden layer and the output layer. The first one possesses \( N_c \) elementary distance neurons (by extension called centres or kernels) and operates a non-linear transformation of the input space. The second one achieves linear combinations of the centre outputs, then activates the general outputs through activation functions.

The partitioning approach leads to a splitting of the complete K-class classifier into K sub-classifiers dedicated to the separation of one class from the others. The figure 2 shows the subnet relating to the class \( \Omega_j \).

\[ \text{Figure 2: RBF subnet structure} \]

In order to estimate the posterior probability class membership, each subnet has a single output, that is activated through a sigmoid function:

\[ \text{Prob}(\Omega_j | X) = F\left(w_0 + \sum_{i=1}^{N_c} w_i \phi\left(\|X - C_i\|\right)\right) \]  \hspace{0.5cm} (4)

with \( j=1..K \) and \( F \) the sigmoid function,

\[ F(x) = \frac{1}{1 + \exp(-x)} \]  \hspace{0.5cm} (5)

The figure 3 presents the complete classifier structure. For the belonging to the classes, the decision rule is very simple and corresponds to the maximum posterior probability.

The partitioning approach allows us to independently design each subnet (different centre number and positions, different widths). The training phase adjusts the weights of the output layer, one subnet after another, and the input vectors can be personalized in order to be more adapted to the separability of each particular class.

\[ \text{Figure 3: RBF complete network structure with 1-by-K partitioning} \]

Other partitioning strategies can be worked out as 2-by-2 partitioning that assigns a subnet to each pair of classes. However, the decision rule becomes more complicated [3].

The learning of RBF networks can be global: the position of centres \( C_i \), the standard deviation of Gaussian kernels \( \beta_i \) and the weights \( w_i \) are simultaneously adjusted with the help of iterative gradient algorithm in order to minimize the least error square cost function [2]. This global learning can present convergence difficulties.

A hybrid learning has been adopted that firstly adjusts the number, position and width of centres, then sets the weights.
4. Number and position of centres

The positioning of the radial basis centres is a crucial problem for the RBF networks. In supervised mode, an easy solution for selecting the centres of the jth subnet is to set them equal to the input vectors of the training set corresponding to the class j. This solution is unrealistic when the training set dimension becomes large.

The K-nearest-neighbour method [4] or the K-means clustering [5] can be used to reduce the number of centres. This article details the selection of centres based on the Orthogonal Least Square technique [6] [7] and its implementation in the case of generalized distance (2).

4.1. Orthogonal least square method

In this technique, the network is considered as a linear regression model of its output. In the case of the first subnet, providing the separation of $\Omega_1$ among the other classes, the output model is equal to:

$$S(X) = \sum_{i=1}^{N_1} w_i \phi(\|X_i - X\|_A) + \varepsilon$$  \hspace{1cm} (6)

$N_1$ is the number of data of the class $\Omega_1$. $N$ is the total number of observations of the training set. If $X$ skims through all the data set, we recognize the equation (4) with linear output activation $F(x) = x$ and without bias. In matrix form:

$$\mathbf{S} = \Phi \mathbf{w} + \mathbf{v}$$  \hspace{1cm} (7)

$\mathbf{w}$, the regression parameters,

$$\Phi = \begin{bmatrix} \Phi_1 & \Phi_2 & \ldots & \Phi_{N_1} \end{bmatrix}$$

$$\begin{bmatrix} \|X_1 - X\|_A \|X_2 - X\|_A \ldots \|X_{N_1} - X\|_A \\ \|X_1 - X_2\|_A \|X_2 - X_2\|_A \ldots \|X_{N_1} - X_{N_1}\|_A \\ \|X_1 - X_{N_1}\|_A \ldots \|X_{N_1} - X_N\|_A \end{bmatrix}$$

The output is set to 0 or 1 according to the class belonging. In the case of the first subnet, $\mathbf{S}$ is equal to,

$$\mathbf{S} = \begin{bmatrix} 1 & 1 & 1 & 0 & \ldots & 0 & 0 & 0 \\ \vdots & \ddots & \ddots & \vdots \end{bmatrix}$$

The Orthogonal Least Square algorithm (OLS) is a sequential method, adapted from the Gram-Schmidt orthogonalization procedure. The algorithm starts by considering all $N_1$ data of the class $\Omega_1$ as centres, then ranks them from the most to the least relevant centres: at the first step, the relevance of each centre is estimated by measuring the angle between $\mathbf{S}$ and each column vector $\Phi_i$ of $\Phi$. The column $i$ (corresponding to the centre $X_i$) with the lowest angle is ranked first. The remaining vectors $\Phi_{k+1}$, and the output, are orthogonalized with respect to the column vector $\Phi_i$. At the second iteration, the remaining centres are ranked again by estimating the relevance in a new $N-1$ dimension space; the procedure is repeated until all centres are ranked (cf. figure 4). The OLS algorithm must be carried out separately for each subnet.

![Figure 4: OLS algorithm for the selection of the centres of the first subnet](image)

The resolution of the regression problem leads to a solution for the weights $\mathbf{w}$. This solution will not be used in the generalization phase because the output activation function was linear during the choice of centres. Sigmoid functions are preferred for pattern classification and an iterative delta-generalized rule will be used for the estimation of weights.

In spite of the non linearity of the output layer, the using of a linear regression model will be justified a posteriori by obtaining good classification performances (cf. §6) : in our particular study, the non linearities of the output layer do not have a lot of influence on the centre relevances.

4.2. Usual stopping criterion

For the choice of centre number, a feature selection stopping criterion [6] can be applied. It consists of an iterative evaluation of the contribution of each ranked centre to the output given by,
\[
oi^2 = \frac{\Phi_i^{(k-1)'}}{\Phi_j^{(k-1)'}} \sum_{k=1}^{\kappa} f_i^{(k-1)} \Phi_j^{(k-1)} \tag{8}\]

During the OLS iterations, the contributions of the already ranked centres can be summed (because of the Gram-Schmidt Orthogonalization) and normalized:

\[
\rho_j = \frac{\sum_{i=1}^{\kappa} o_{ij}^2}{\sum_{j=1}^{\kappa} \sum_{i=1}^{\kappa} o_{ij}^2} \tag{9}\]

The final number of centres is obtained when \( \rho_j \) reaches a target value between 0 and 1.

### 4.3. Stopping criterion with random centre addition.

This section introduces an other stopping criterion based on the addition of a random center [8] [9] that provides an intuitive –but principled– approach to the selection procedure.

Starting with all \( N_1 \) data of the class \( \Omega \) as candidate centres, the idea is to add to the matrix \( \Phi \), a new column corresponding to a random centre. Then, the OLS procedure is applied to order the \( N_1 + 1 \) centres and the centres ranked after the random one will be considered as irrelevant and can be removed.

In practice, the rank of the random centre is also a random variable. In order to evaluate correctly the relevancy of each candidate centre, a large number of random centres must be generated: the OLS procedure is operated each time in order to obtain an estimation of the probability density function \( f \) of the rank of the random centre, then its cumulative distribution function \( F \) (cf. figure 5).

The using of a significance level \( \alpha \) allows us to determine the final number of centres by means of the equation,

\[
F(N_c) = \alpha \tag{10}\]

So, the probability that a random centre is more relevant than one of the \( N_c \) centres ranked first is \( \alpha\% \).

### 5. Choice of distance type and centre width

In RBF networks, the width parameter \( \beta \) controls the overlapping of each radial basis function. If \( \beta \) is small, a high number of centres will be required for the training set description; there is over-parametrization. If \( \beta \) is large, the centre includes too extensive an area of the representation space. In both cases, the generalization capabilities of the network will be poor. Obviously, the standard deviation \( \beta \) of each neuron must be set in close correlation with the distance type (Euclidian distance, Mahalanobis distance...) induced by the choice of normalization matrix \( A \) (eq. 3). Four couples width-distance have been tested. The next table lists them.

| Distance   | Width                  | \( N_c \) : subnet centre number | \( d \) : dimension of subnet input vector | \( A \) :
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Euclidian</td>
<td>( \beta = \left| \xi_i - \xi_j \right|_A )</td>
<td>( \left( \begin{array}{ccc} 1 &amp; \ldots &amp; 0 \ \vdots &amp; \ddots &amp; \vdots \ 0 &amp; \ldots &amp; 1 \end{array} \right) )</td>
<td>( \left( \begin{array}{ccc} \sigma_1^2 &amp; \ldots &amp; 0 \ \vdots &amp; \ddots &amp; \vdots \ 0 &amp; \ldots &amp; \sigma_d^2 \end{array} \right) )</td>
<td>( \left( \begin{array}{ccc} \sigma_{11}^2 &amp; \ldots &amp; \sigma_{1d}^2 \ \vdots &amp; \ddots &amp; \vdots \ \sigma_{dd}^2 &amp; \ldots &amp; \sigma_{dd}^2 \end{array} \right) )</td>
</tr>
<tr>
<td>Mahalanobis</td>
<td>( \beta = \left| \xi_i - \xi_j \right|_A )</td>
<td>( \left( \begin{array}{ccc} \sigma_1^2 &amp; \ldots &amp; 0 \ \vdots &amp; \ddots &amp; \vdots \ 0 &amp; \ldots &amp; \sigma_d^2 \end{array} \right) )</td>
<td>( \left( \begin{array}{ccc} \sigma_{11}^2 &amp; \ldots &amp; \sigma_{1d}^2 \ \vdots &amp; \ddots &amp; \vdots \ \sigma_{dd}^2 &amp; \ldots &amp; \sigma_{dd}^2 \end{array} \right) )</td>
<td>( \left( \begin{array}{ccc} \sigma_{11}^2 &amp; \ldots &amp; \sigma_{1d}^2 \ \vdots &amp; \ddots &amp; \vdots \ \sigma_{dd}^2 &amp; \ldots &amp; \sigma_{dd}^2 \end{array} \right) )</td>
</tr>
</tbody>
</table>

| Table 1 : different couples of Distance-Width |
In the first and second case, the distance is the Euclidian distance and the width $\beta$ is deduced from heuristic rules; in case 1, $\beta$ is equal to the average distance between the subnet centres; in case 2, $\beta$ is linked to the maximum distance between the subnet centres [2]; the kernels' own radial symmetry.

In the third and fourth case, the kernels are no longer symmetric; all components are separately processed by using a specific matrix $A$. In case 3, the $j$th component width is normalized by the $j$th variance. In case 4, the correlation between components are taken into account and the normalization is operated by means of the covariance matrix. The distance is referred to as Mahalanobis distance [10]. The additional parameter $\beta$ is set to the average $A$-distance between the subnet centres.

If the reduction of kernel number has been efficient, the covariance matrix could be estimated from too small a number of centres and its inversion can present some numerical difficulties. One solution is to estimate the covariance matrix from the complete training subset relating to the treated class.

6. Application to classification of defect signatures in Non Destructive Evaluation domain

The application concerns the classification of rail head defect signatures measured by a specific eddy current sensor [11]. After a parametrization of the raw signals, a classification procedure is carried out in order to assign each defect to one of the four defined classes of defects (cracks, welded joints, shellings, joints before crossing). The training set is made of 140 experimental data (64 examples from the class 1, 39 from the class 2, 16 from the class 3 and 21 from the class 4).

The classifier is partitionned into four RBF subnets. For the centre selection of each subnet, the OLS procedure is used to rank the training centres. The centre number is given by the stopping criterion above mentioned (9). The figure 6 shows the cumulative distribution function of the random centre rank relative to the first subnet.

With a target significance level of 0.05, the four subnets are constituted with respectively 5, 5, 3 and 5 centres each, in accordance with the designed structure (figure 3); the dimension of the input vectors are 15, 15, 8 and 9 and they have been adjusted by using the same OLS approach as the centre selection one (feature selection with random feature addition [8]).

The choice of distance is illustrated by figure 7 which compares the four couples' distance-width (table 1). For each distance case, the output $S$ of a single kernel of the subnet n°4 is presented while the input $X$ scans the whole training set. In figure 7, the classes are delimited by vertical lines.

In case 1, $\beta$ is quite obviously too large and the kernel includes almost all the training set. The cases 2 and 3 are more efficient, but the separability of class 4 is not guaranteed. The case
4 (Mahalanobis distance) allows the best separability of class 4. In fact, it is always advisable to abandon Euclidian distance when the course of input vector components are highly different.

The global classification performances for our application are presented table 2. A "leave one out" procedure applied to the training set, allows us to obtain these results with the proposed structure of the classifier.

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>case 1</td>
<td>81%</td>
</tr>
<tr>
<td>case 2</td>
<td>79%</td>
</tr>
<tr>
<td>case 3</td>
<td>88%</td>
</tr>
<tr>
<td>case 4</td>
<td>97%</td>
</tr>
</tbody>
</table>

Table 2 : well-classification performances

The well-classification performances reach 97% with the Mahalanobis distance.

6. Conclusion

This article has presented a partitioning approach for the classification problem using RBF networks. The global K-class problem is split into K sub problems dedicated to the separation of one class from the others ; this makes it possible to optimize the subnets independently in terms of complexity.

The Orthogonal Least Square method provides a suitable solution for the choice of centres of RBF networks among a candidate set of learning data. The ranking of the centres is carried out by taking into account the class separability unlike a random selection or even a k-means clustering. The choice of the centre number is deduced from an intuitive stopping criterion using the addition of a random centre ; this method allows the kernel number of the input layer to be efficiently reduced.

The training of the networks has been illustrated by a particular Non Destructive Evaluation problem involving the discrimination of four classes of defects.

For this application, classification performances are given for Euclidian and Mahalanobis distances. The Mahalanobis distance turns out to be more effective for the processing of the classification problem, because of the different order of magnitude of the data component variations.

References :

[1]: Bishop C.M. Neural Networks for Pattern Recognition. Clarendon Press 1995
[7]: Chen S. Cowan S.F. Grant P.M. Orthogonal Least Square learning algorithm for radial basis function networks. Neural Networks vol 2 n°2 pp.302-309. 1991