DESIGNING RBF CLASSIFIERS FOR WEIGHTED BOOSTING

Vanessa Gómez-Verdejo, Jerónimo Arenas-García,
Manuel Ortega-Moral and Aníbal R. Figueiras-Vidal
Department of Signal Theory and Communications

Universidad Carlos III de Madrid
Avda. Universidad 30, 28911, Leganés (Madrid), SPAIN.
{*vanessa,jarenas,ortegam,arfv*}*@tsc.uc3m.es*

***Abstract*— The recent interest in combining Neural Networks has produced a variety of techniques. This paper deals with boosting methods, in particular, Real AdaBoost schemes built up with Radial Basis Function Networks. Real Adaboost emphasis function can be divided into two different terms, the first only focus on the quadratic error of each pattern and the second only takes into account its "proximity" to the boundary. Incorporating to this fixed emphasis function an additional degree of freedom, that allows us to weight these two terms, and also to select the Radial Basis Functions centroids according to the emphasized regions, we show performance improvements: an error rate reduction, a faster convergence, and overfitting robustness.**

I. INTRODUCTION

Multi-net systems are a good approach to solve difficult tasks which usually require a very complex net, overcoming sizing and training difficulties. Consequently, during the last years there has been an intensive research work to design Neural Networks (NN) ensembles, following different approaches [14]. Among all these procedures, boosting schemes [12], and in particular Real Adaboost (RA), [2] have demonstrated excellent performance.

Boosting designs have their roots in the consideration of weak learners in the light of the Probability Approximately Correct Learning Theory [15]. In fact, these designs have become very popular as a way to obtain advantage of "weak" learners. Concretely, RA works by adding sequentially a new base learner trained with an emphasized population, mainly paying its attention on the most erroneous samples (a detailed description can be found in [13]).

During the last years, many researchers have studied RA convergence properties, proposing different alternatives to improve its performance, making the list of applications where boosting methods are employed grow rapidly [8]. Some of us have also chosen this research line, concretely, we have studied how RA emphasis works, getting some interesting results. For instance, Arenas-García et al. showed in [6] that, although in an indirect manner, these schemes end by concentrating their attention on the samples near the classification boundary. Consequently, in [5] we analyzed how the performance of these schemes can be improved by focusing directly on the samples near the boundary. Furthermore, in this last work, we also studied in detail the RA emphasis function, showing that this function is made up of two different terms which are combined in a fixed way: the first pays attention to the quadratic error of each pattern; the other takes into account its "proximity" to the boundary. In [4] we show that an alternative emphasis function adding a variable mixing parameter allows to combine both terms in a more flexible manner and provides some advantages. We showed, by studying the performance of the new approach in Multi Layer Perceptron (MLP) based RA scheme, that usually the best way to resample the population is emphasizing neither most erroneous samples nor boundary ones, but a particular tradeoff between them.

In this paper we go deeper into this idea; first, we show that similar advantages to those reported when using MLPs can be obtained when boosting Radial Basis Function Networks (RBFN). Second, to maximize the weighting emphasis benefit, we propose a method to appropriately design the Radial Basis Functions (RBFs) of the base classifiers according to the emphasized regions. As we check in several experiments, selecting RBF centroids according to the kind of emphasis that is needed to get these ensembles offer a good performance.

The rest of the paper is organized as follows: first, the classical RA algorithm will be described, paying special attention to its emphasis function; in Sections III and IV we present the weighted emphasis function and the proposed algorithm to design RBFN for boosted ensembles, respectively. Section V is devoted to check the validity of these approaches in some benchmark problems. Finally, in Section VI, conclusions and future research lines will be presented.

II. REAL ADABOOST

The fundamental idea of RA is to combine several "weak" learners in such a way that the ensemble improves their performance. To build up an RA classifier, at each round $t = 1, ..., T$ a new base learner is added implementing a function $o_t(\mathbf{x}_i) : X \longrightarrow [-1, 1]$ aimed to minimize the following error function

$$E_t^2 = \sum_{i=1}^{L} D_t(i)(d_i - o_t(\mathbf{x}_i))^2 \quad (1)$$

where L is the number of training patterns, $d_i \in \{-1, 1\}$ is the target for pattern $\mathbf{x}_i$, $o_t(\mathbf{x}_i)$ is the "weak" learner output

for $\mathbf{x}_i$, and $D_t(i)$ is the weight that the t-th learner emphasis function assigns to $\mathbf{x}_i$. Initially, all weights have the same value $D_1(i) = 1/L$, $\forall i = 1, ..., L$, and they are then updated according to

$$D_{t+1}(i) = \frac{D_t(i)\exp(-\alpha_t o_t(\mathbf{x}_i)d_i)}{Z_t} \tag{2}$$

where Z_t is a normalization factor assuring that $\sum_{i=1}^{L} D_t(i) = 1$, and α_t is the weight assigned to the t-th weak learner. The overall output of the net, $f_T(\mathbf{x})$, is calculated as the weigthed combination of all learners:

$$f_T(\mathbf{x}) = \sum_{t=1}^{T} \alpha_t o_t(\mathbf{x}) \tag{3}$$

weights α_t are calculated at each round according to

$$\alpha_t = \frac{1}{2}\ln\left(\frac{1+r_t}{1-r_t}\right) \tag{4}$$

where r_t, typically named edge of the the t-th learner, is

$$r_t = \sum_{i=1}^{L} D_t(i) o_t(\mathbf{x}_i) t_i \tag{5}$$

This choice of α_t values assures that the following training error bound is minimized

$$E_{\text{train}} = \sum_{i=1}^{L} \mid \text{sign}(f(\mathbf{x}_i)) \neq t_i \mid \leq \sum_{i=1}^{L} \exp(-f_t(\mathbf{x}_i)t_i) \tag{6}$$

Furthermore, as it was demonstrated in [9] , this same criterion maximizes the classification margin defined as

$$\rho = \min_{i=1...L} f_t(\mathbf{x}_i)d_i \tag{7}$$

As explained [5], (2) can be alternatively expressed as

$$D_{t+1}(\mathbf{x}_i) = \frac{1}{Z'_t}\exp\left(\frac{(f_t(\mathbf{x}_i) - d_i)^2}{2}\right)\exp\left(-\frac{f_t^2(\mathbf{x}_i)}{2}\right) \tag{8}$$

where each factor stands for a different kind of emphasis:

$$\text{Error emphasis} \qquad \exp\left(\frac{(f_t(\mathbf{x}_i) - t_i)^2}{2}\right) \tag{9}$$

$$\text{Boundary emphasis} \qquad \exp\left(-\frac{f_t^2(\mathbf{x}_i)}{2}\right) \tag{10}$$

As summary, RA emphasis function (2) can be seen as fixed combination of two terms, one that pays attention to the quadratic error of each pattern (9), and other that takes into account its proximity to the boundary (10).

III. A WEIGHTED EMPHASIS FUNCTION

In the light of (8), one may wonder if this fixed combination of emphasis terms is optimal in all situations. In fact, as we have already checked in [4], different combinations of these terms can improve the ensemble performance in general situation. There, we proposed to combine (9) and (10) by means of a weighting parameter λ ($0 \leq \lambda \leq 1$),

$$D_{\lambda,t+1}(i) = \frac{1}{Z_t}\exp\left(\lambda \cdot (f_t(\mathbf{x}_i) - t_i)^2 - (1-\lambda)\cdot f_t^2(\mathbf{x}_i)\right) \tag{11}$$

This more flexible combination allows to pay more or less attention to the boundary "proximity" or to the quadratic error of each sample by selecting different values λ. Three special cases, corresponding to different values λ, should be remarked:

i) $\lambda = 0$: only "proximity" to the boundary is taken into account.

$$D_{\lambda=0,t+1}(i) = \frac{1}{Z_t}\exp\left(-f_t^2(\mathbf{x}_i)\right) \tag{12}$$

ii) $\lambda = 0.5$: the classical RA emphasis function (2) is used.

iii) $\lambda = 1$: quadratic error only is considered.

$$D_{\lambda=1,t+1}(i) = \frac{1}{Z_t}\exp\left((f_t(\mathbf{x}_i) - t_i)^2\right) \tag{13}$$

In this paper we study the effects of using this weighted emphasis function for the design of a boosted ensemble built up by local classifiers, concretely, RBFNs. In [4] we showed that a good selection of the weighting parameter can reduce the error rate, accelerate the convergence and, even, avoid the overfitting problem, when building up MLP base multi-net systems. However, when the weak learners are RBFNs we have to design them carefully in order to make the most of the weighted emphasis; as we realized in [5], studying the first two particular cases[1], it is necessary to select the centroids and to adjust its dispersion parameter according to the kind of emphasis. So, in order to solve this problem and, therefore, improve the performance of RBFN base boosted ensembles, in the next section we present a new method to design the RBFs according the emphasized regions.

IV. BUILDING RBFNS FROM EMPHASIZED DATASET

An RBF network consists on a group of K basis functions linearly combined to get the network output

$$o_t(\mathbf{x}) = \sum_{k=1}^{K} w_k g_k(\mathbf{x}) \tag{14}$$

where $g_k(\cdot)$ is a function with circular symmetry. We will typically use Gaussian kernels with centers $\mathbf{c}_k$ and dispersion parameters β_k

$$g_k(\mathbf{x}) = \exp\left(-\frac{\parallel \mathbf{x} - \mathbf{c}_k \parallel^2}{2\beta_k^2}\right) \tag{15}$$

When we build up a boosted ensemble by combining RBFNs, we train their output weights, w_k, in order to minimize (1), but we have also to design the input layer, i.e.,

[1]We have tested a normalized version of (11) when $\lambda = 0$ and $\lambda = 0.5$.

we have to select centroids $\mathbf{c}_k$ and dispersion parameters β_k. As we have already pointed out, it is quite adequate selecting these parameters according to the emphasized regions. So, to design radial functions $g_k(\cdot)$ we propose to select both centers and dispersion parameters according to probability distribution D_{centers}; for the case of binary classification, this method can be summarized as follows:

1) We use D_{centers} distribution to independetly select K_{-1} centroids from the samples corresponding to class C_{-1}, and K_1 centroids from those with positive targets, where K_{-1} and K_1 are chosen according to the a priory probability of each class.
2) Now, for each centroid, we calculate the Euclidean distances, weighted according to D_{centers}, from centers $\mathbf{c}_k$ to all the patterns attributed to that centroid,

$$\text{dist}(\mathbf{x}_i, \mathbf{c}_k) = N_{c_k} D_{\text{centers},k}(\mathbf{x}_i) \parallel \mathbf{c}_k - \mathbf{x}_i \parallel \qquad (16)$$

where

$$D_{\text{centers},k}(\mathbf{x}_i) = \frac{D_{\text{centers}}(\mathbf{x}_i)}{\sum_{\mathbf{x}_i \in C_k} D_{\text{centers}}(\mathbf{x}_i)} \qquad (17)$$

C_k is the set of all patterns belonging to the centroid c_k, and N_{c_k} is the number of patterns in C_k.
3) Finally we calculate the dispersion parameter of each centroid, β_k, as

$$\beta_k = \frac{\mu_k^2}{\sigma_k} \qquad (18)$$

where μ_k and σ_k are, respectively, the mean and standard deviation of $\text{dist}(\mathbf{x}_\text{i}, \mathbf{c}_\text{k})$, calculated as

$$\mu_k = \frac{1}{N_{c_k}} \sum_{\mathbf{x}_i \in C_k} \text{dist}(\mathbf{x}_i, \mathbf{c}_k) \qquad (19)$$

$$\sigma_k = \sqrt{\frac{1}{N_{c_k}} \sum_{\mathbf{x}_i \in C_k} \left(\text{dist}(\mathbf{x}_i, \mathbf{c}_k) - \mu_k\right)^2} \qquad (20)$$

V. Experiments

To show the performance of combining a weighted emphasis with an adequate RBF design, we have built a series of ensembles and we have evaluated them when applied to several binary problems. In particular, we have selected four binary problems from [1]: *Abalone* (a multiclass problem converted to binary according to [11]), *Breast*, *Image* and *Ionosphere*; and two synthetic problems: *Kwok* [7] and *Ripley* [10]. In Table I we have summarized their main features: number of dimensions (dim) and number of samples of each class $\{C_1/C_{-1}\}$ in the training and test set. Some of the problems have a predefined test set; when this was not the case, ten random partitions with 40% of the data set have been selected to test the performance of the classifier.

Each of these ensembles is a linear combination of RBFs with input layer designed according to the method proposed in Section IV, and output weights trained to minimize cost function (1) by means of a sthocastic gradient descent algorithm with a learning step linearly decreasing from 0.1 to 0 along 50 epochs [2]. The selection of ensemble output weights α_t is done according to (4); in this way, we are still minimizing a bound on the training error, and simultaneously maximizing the classifier margin, as RA does.

TABLE I

Main features of the benchmark problems.

Problem	dim	# Train samples	# Test samples
Abalone	8	1238/1269	843/827
Breast	9	275/145	183/96
Image	18	821/1027	169/293
Ion	34	275/145	183/96
Kwok	2	300/200	6120/4080
Ripley	2	125/125	500/500

To stand out the importance of an optimal tradeoff between error and boundary emphasis, we have built different ensembles emphasizing the RBF design and the output layer training according to (11), where λ has been explored from 0 to 1 with a step of 0.1.

Table II displays the mean value and, between brackets, the standard deviation of test errors averaged over 50 independent runs, for classical RA ($\lambda = 0.5$) and the weighting parameter that achieves the best result, denoted as λ_0. For each binary problem, we have used a different number of rounds T, in order to assure a complete convergence of the ensemble, and RBFs with different representational power, concretely, 2% and 10% of the number of training data have been selected as centroids (denoted as $\%K$ in the Tables). Furthermore, to measure the statistical importance of the different approaches, we have used the Wilcoxon Rank Test (WRT) [3], where a value lower than 0.1 indicates that the differences between them are significant[3]; on the contrary, this value is close to 1 when there is no statistical difference between the two error rates.

It can be seen that the fixed combination of the emphasis terms that RA employs is not always the best tradeoff. In some problems emphasis focused mainly on the boundary patterns ($\lambda < 0.3$) achive a significant error rate reduction, for instance, in *Abalone*, *Kwok*, or in *Image* when the percentange of centers of the weak learners is 2%. In other problems, like *Ripley*, it is better to concentrate the emphasis on the most erroneous patterns. An optimal tradeoff between the emphasis terms not only reduces the error rate, but also can avoid the overfitting problem that RA presents and it can accelerate the algorithm convergence. For instance, in *Ionosphere* (see Figure 1), we get a higher error rate reduction for large λ values, while lower λ values offer lower error rates.

In order to show the improvement achieved boosting the center selection, we present in Table III the error rates obtained if we only emphasize the population to train the output layer

[2]This number of epochs ensures the network convergence.
[3]Values lower than 0.01 have been rounded down to 0.

TABLE II

TEST ERRORS DESIGNING RBFS CENTROIDS WITH EMPHASIS

Prob.	% K	T	E_{RA}	λ_0	E_{λ_0}	WRT
Abal	10	100	20.01 (0.23)	**0**	**19.89 (0.27)**	0.03
	2	400	20.49 (0.20)	**0**	**20.32 (0.25)**	0
Brea	10	100	3.47 (0.93)	**0.1**	**3.18 (0.74)**	0.17
	2	200	3.49 (0.85)	**0.3**	**3.41 (0.88)**	0.49
Imag	10	200	**3.18 (0.21)**	0.5	3.18 (0.21)	1
	2	200	3.93 (0.32)	**0.1**	**3.79 (0.21)**	0.01
Ion	10	100	**1.36 (0.50)**	0.5	1.36 (0.50)	1
	2	200	2.57 (0.23)	**0.2**	**2.16 (0.32)**	0
Kwok	10	100	12.87 (0.33)	**0**	**11.97 (0.17)**	0
	2	400	12.18 (0.14)	**0.2**	**11.83 (0.11)**	0
Rip	10	100	9.53 (0.34)	**0.8**	**8.97 (0.2)**	0
	2	200	9.01 (0.17)	**0.7**	**8.80 (0.2)**	0

TABLE III

TEST ERRORS DESIGNING RBFS CENTROIDS WITHOUT EMPHASIS

Prob.	% K	T	E_{RA}	λ_0	E_{λ_0}	WRT
Abal	10	100	19.84 (0.39)	**0.7**	**19.81 (0.41)**	0.46
	2	400	20.64 (0.37)	**0.6**	**20.58 (0.34)**	0.31
Brea	10	100	3.50 (0.85)	**0.3**	**3.22 (0.82)**	0.33
	2	200	3.69 (0.76)	**0.3**	**3.54 (0.83)**	0.41
Imag	10	200	4.31 (0.48)	**0.4**	**4.16 (0.54)**	0.08
	2	400	5.18 (0.37)	**0.3**	**4.86 (0.32)**	0
Ion	10	100	3.48 (1.54)	**0.1**	**1.81 (0.40)**	0
	2	400	**2.76 (0.73)**	0.5	2.76 (0.73)	1
Kwok	10	100	13.07 (0.29)	**0**	**12.09 (0.19)**	0
	2	200	11.90 (0.13)	**0.4**	**11.81 (0.09)**	0
Rip	10	100	10.35 (0.49)	**0.9**	**9.24 (0.26)**	0
	2	400	9.36 (0.21)	**0.6**	**9.18 (0.25)**	0

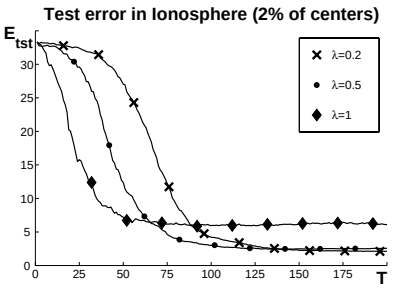

Fig. 1. Evolution of the test error for λ 0.2, 0.5 and 1

of the RBFNs; i.e., we have designed the RBF according to the method presented in Section IV, but, now, probability distribution D_{centers} is always a uniform distribution, independent from the output layer emphasis function that remains applied according to (11).

The proposed method offers for every problem but *Abalone* an error rate improvement, reducing both its mean value and its standard deviation. Furthermore, in problems like *Ripley*, *Image* or *Ionosphere*, a faster convergence can be observed.

Finally, trying to improve even more the boosted ensemble performance, we have separated the weighting parameter of the output layer from that used for the centroid design; i.e., we apply the emphasis function (11) in both training procedures, but different weighting parameters: λ_{out} for training the RBFN output layer and λ_{cen} for emphasizing the centroids selection.

This approach has been tested in two particular problems, *Abalone* and *Kwok*, when the RBFNs have as centroids 10% of the training data. In particular, several ensembles have been built varying weighting parameters λ_{out} and λ_{cen} in range $[0, 1]$ with a step of 0.1; the best error rates of this approach are shown in Table IV.

Finally, we can see in Figure 2, where the error test in the problem *Kwok* according to λ_{out} and λ_{cen} is depicted,

TABLE IV

TEST ERRORS SEPARATING THE RBFN LAYER EMPHASIS

Probl.	% K	T	E_{RA}	λ_{out}	λ_{cen}	$E_{\lambda_{out},\lambda_{cen}}$
Abal.	10	100	20.01 (0.23)	**0.5**	**0.3**	**19.69 (0.31)**
Kwok	10	100	12.87 (0.33)	**0.2**	**0.9**	**11.92 (0.13)**

how by tuning these parameters we can modify considerably the ensemble performance. For example, when the centroids design is boosted focusing on the most erroneous patterns, $\lambda_{cen} > 0.9$, the error rate is about 12%, and λ_{out} value introduces slight changes; whereas, when the centroids are placed on the boundary, λ_{out} value is critical, and it can change the error rate from 12% to 13.45%. Of course, it is necessary a good tradeoff between these parameters to get the best results: selecting the centroids according to the most erroneous patterns ($\lambda_{cen} = 0.9$) and focusing the training of the RBF output layer on the patterns nearest to the boundary ($\lambda_{out} = 0.2$), we get an error rate of 11.92%.

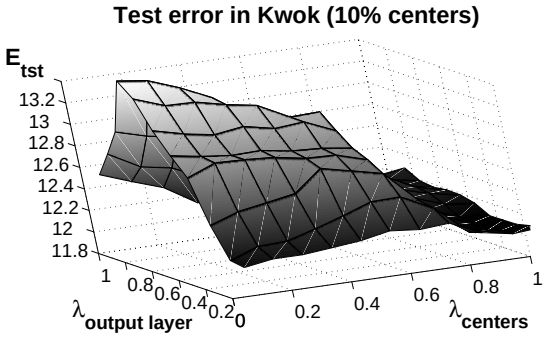

Fig. 2. Test error with regard to λ_{cen} and λ_{out}

VI. Conclusions and future work

In this paper, we check that using different tradeoffs between error and boundary emphasis in boosting methods provides performance improvements. We have specially pointed out the influence of this new boosting strategy on the design of multi-net systems, when local classifiers, such as RBFNs, are used as base learners. In this case, to get the most from emphasis methods it is necessary to apply modified criteria for selecting the centroids and dispersion parameters of the RBFs.

Experiments show, in several benchmark problems, that a good selection of the weighting parameter together with a good design of the basis functions, can reduce the error rate, accelerate the convergence of the ensemble, and avoid the overfitting problem of RA schemes. Finally, we have shown how these results can be enhanced even more if the emphasis applied to the RBF output layer training is separated from the emphasis used for designing the basis functions.

These evidences suggest the appropriateness of designing automatic methods to select the optimal value of the weighting parameters. Even more, it would be very interesting to adapt weighting values along the ensemble growing in order to select in the first rounds λ values that get a faster convergence, and in the last rounds λ values that allow a lower error rate. These ideas constitute a promising research line where we are currently working.

Acknowledgment

This work has been partly supported by CICYT grant TIC2002-03713. The work of V. Gómez-Verdejo was also supported by the Education Department of the Madrid Community by a scholarship.

References

[1] C. L. Blake and C. J. Merz. UCI repository of machine learning databases. http://www.ics.uci.edu/ mlearn/MLRepository.html, 1998. University of California, Irvine, Dept. of Information and Computer Sciences.

[2] Y. Freud and R. E. Schapire. Experiments with a new boosting algorithm. In *Proc. 13th Intl. Conf. Machine Learning*, pages 148–156, Bari, Italy, 1996.

[3] J. D. Gibbons. *Nonparametric Statistical Inference*. Basel : Marcel Dekker, New York, 4th edition, 2003.

[4] V. Gómez-Verdejo, M. Ortega-Moral, J. Arenas-García, and A. Figueiras-Vidal. Boosting by weighting boundary and erroneous samples. In *To be presented at ESSAN'05*, Bruges, Belgium, April 2005.

[5] V. Gómez-Verdejo, M. Ortega-Moral, J. P. Cabrera, J. Arenas-García, and A. Figueiras-Vidal. Boosting by emphasizing boundary samples. In *Proc. of the Learning'04 Intl. Conf.*, pages 67–72, Elche, Spain, 2004.

[6] A. J. C. Sharkey J. Arenas-García, A. R. Figueiras-Vidal. The beneficial effects of using multi net systems that focus on hard patterns. In F. Rolli T. Windeatt, editor, *Multi Classifier Systems (Proc. 4th Intl. Workshop)*, pages 45–54, Survey, U.K., 2003. Springer-Verlag (LNCS).

[7] J. T. Kwok. Moderating the output of support vector classifiers. *IEEE Trans. on Neural Networks*, 10(5):1018–1031, 1999.

[8] R. Meir and G. Ratsch. An introduction to boosting and leveraging. In S. Mendelson and A. Smola, editors, *Advanced Lectures on Machine Learning*, LNCS, pages 119–184. Springer Verlag, 2003.

[9] G. Ratsch., T. Onoda, and K. R. Muller. Soft margins for AdaBoost. *Machine Learning*, 42(3):287–320, March 2001.

[10] B. D. Ripley. Neural networks and related methods for classification (with discussion). *J. Roy. Statist. Soc. Series*, 56:409–456, 1994.

[11] A. Ruiz and P. E. López de Teruel. Nonlinear kernels-based statistical pattern analysis. *IEEE Trans. on Neural Networks*, 12(1):16–32, 2001.

[12] R. E. Schapire. The strength of weak learnability. In *30th Annual Symposium on Foundations of Computer Science*, pages 28–33, 1989.

[13] R. E. Schapire and Y. Singer. Improved boosting algorithms using confidence-rated predictions. *Machine Learning*, 37(3):297–336, 1999.

[14] A. J. C. Sharkey. *Combining Artificial Neural Nets. Ensemble and Modular Multi-Net Systems*. Springer-Verlag, London, UK, 1999.

[15] L. Valiant. Theory of the learnable. *Communications of the ACM*, 27:1134–1142, 1984.