ESANN'2006 proceedings - European Symposium on Artificial Neural Networks
Bruges (Belgium), 26-28 April 2006, d-side publi., ISBN 2-930307-06-4.

Designing neural network committees by combining boosting ensembles

Vanessa Gómez-Verdejo and Aníbal R. Figueiras-Vidal *

Department of Signal Theory and Communications, Universidad Carlos III de Madrid
Avda. Universidad 30, 28911 Leganés (Madrid), SPAIN.
{*vanessa, arfv*}*@tsc.uc3m.es*

Abstract.
The use of modified Real Adaboost ensembles by applying weighted emphasis on erroneous and critical (near the classification boundary) has been shown to lead to improved designs, both in performance and in ensemble sizes. In this paper, we propose to take advantage of the diversity among different weighted combination to build committees of modified Real Adaboost designs. Experiments show that the expected improvements are obtained.

1 Introduction

To combine Neural Networks (NNs) is more effective than trying to solve difficult problems by using big size single NNs; not only an easier design and better accuracy can be obtained, but also a more clear understanding of how the resulting machine works. These reasons have increased the interest in this research area during the recent years, producing a wide variety of methods [9]. Among them, specially boosting methods [7], and in particular Real Adaboost (RA) [2, 8], are attractive because their conceptual principles and their proved good performance.

In [3] and [4] we proposed a new weighted emphasis function that allows, introducing a mixing parameter λ, to assign more or less importance to the most erroneous or to the critical (those that lie close to the classification boundary) patterns; we call this procedure the RA with Weighted Emphasis (RA-WE) algorithm. Some early experiments showed us that significant improvements over the classical RA performance can be achieved when the mixing parameter λ is adequately selected. However, finding the optimal λ is not an easy task, and despite it can be found using cross-validation, it is still a delicate issue [5].

Rather than trying to find a good value for λ, in this paper we propose to combine the outputs of a number of RA-WE networks trained with different values of λ. This way, we can take advantage of the diversity among all RA-WE components. In fact, it will turn out that an appropriate combination can even improve the accuracy of the best RA-WE element due to a reduction in the error variance and, therefore, a better generalization capabilities [9].

*This work has been partly supported by MEC grant TEC 2005/00992. The work of V. Gómez-Verdejo was also supported by the Chamber of Madrid Community and European Social Fund by a scholarship.

The rest of the paper is organized as follows: in the next section, a brief description of the RA-WE algorithm is presented; later, we will consider the issue of how to appropriately combine a number of these RA-WE networks trained with different settings; in Section 4 we will test the performance of the resulting schemes over several benchmark problems. Finally, conclusions and future research lines will close the paper.

2 The Real Adaboost with Weighted Emphasis Algorithm

As for conventional Real Adaboost [8], the construction of a RA-WE network is carried out by incrementally adding to the ensemble, at each round $t = 1, ..., T$, a new base learner implementing a function $o_t(\mathbf{x}) : \mathcal{X} \longrightarrow [-1, 1]$. The overall ensemble output at round T, $f_T(\cdot)$, is calculated as a linear combination of all learners outputs,

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

where α_t is the output weight assigned the t-th learner, that is calculated according to

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

where

$$r_t = \frac{\sum_{i=1}^{L} \exp[-f_t(\mathbf{x}_i) d_i] o_t(\mathbf{x}_i) d_i}{\sum_{i=1}^{L} \exp[-f_t(\mathbf{x}_i) d_i]} \tag{3}$$

($\{\mathbf{x}_i \in \chi, i = 1, ..., L\}$ being the training data set and $d_i \in \{-1, 1\}$ being the target for pattern $\mathbf{x}_i$), if we accept that the training error bound

$$\sum_{i=1}^{L} \exp[-f_t(\mathbf{x}_i) d_i] \geq E_{\text{train}} = \sum_{i=1}^{L} | \operatorname{sign}[f(\mathbf{x}_i)] \neq d_i | \tag{4}$$

is minimized.

In RA-WE, each learner is trained to minimize the weighted mean square error over the training data set

$$C_t = \sum_{i=1}^{L} D_{\lambda,t}(i) [d_i - o_t(\mathbf{x}_i)]^2 \tag{5}$$

where $D_{\lambda,t}$ is the following mixed emphasis function

$$D_{\lambda,t+1}(i) = \frac{1}{Z_t} \exp \left\{ \lambda [f_t(\mathbf{x}_i) - d_i]^2 + (1 - \lambda) f_t^2(\mathbf{x}_i) \right\} \tag{6}$$

, and Z_t being a normalization factor that assures $\sum_{i=1}^{L} D_{\lambda,t+1}(i) = 1$.

This weighted emphasis function allows us, by selecting different values for λ ($0 \leq \lambda \leq 1$), to choose how much attention should be placed on the critical

patterns (those near the classification boundary) and how much on the most erroneous examples. For instance, when $\lambda = 0$ we only focus on critical patterns, those close to the classification boundary ($f_t(\mathbf{x}_i) = 0$); when $\lambda = 1$, the emphasis function only pays attention to the quadratic error of each pattern; intermediate values correspond to intermediate situations, $\lambda = 0.5$ providing the classical RA emphasis function (see [5] for a more detailed explanation).

Different selections of λ result in RA-WE networks with different properties. In [5], we explored a cross-validation strategy to select this parameter for getting a minimum generalization error. Here, we follow a different approach consisting on training a number of RA-WE networks, each one with a different value of λ, and combining their outputs. Obviously, this combination can also be carried out in many different manners, but it turns out that the simple schemes described in the following section give satisfactory results.

3 Committees of RA-WE networks

Let us consider we already have a set of N RA-WE networks trained using N different λ values from the set $\{\lambda_0, \ldots, \lambda_{N-1}\}$. Let us also denote by $\bar{f}_{T,n}(\mathbf{x}), n = 0, \ldots, N-1$, the normalized version of the previous network corresponding to λ_n:

$$\bar{f}_{T,n}(\mathbf{x}) = \frac{f_T(\mathbf{x})}{\sum_{t=1}^{T} |\alpha_t|} \tag{7}$$

where $f_T(\mathbf{x})$ is calculated according (1) and $|\cdot|$ means absolute value.

Among the different methods proposed in the literature to combine neural networks, we have chosen to use a weighted linear combination of the outputs of the RA-WE nets, so that it is possible to take into account the relative accuracies of each combined element [9]. We will consider two different outputs for the overall network, depending on which kind of activation function is used:

- **Linear activation**

$$F_{\text{lin}}(\mathbf{x}) = \sum_{n=0}^{N-1} w_n \bar{f}_{T,n}(\mathbf{x}) \tag{8}$$

- **Hyperbolic tangent activation**

$$F_{\text{th}}(\mathbf{x}) = \tanh\left[\sum_{n=0}^{N-1} w_n \bar{f}_{T,n}(\mathbf{x})\right] \tag{9}$$

In both cases, output weights $\mathbf{w} = \{w_0, ...w_{N-1}\}$ will be selected so that the sum-of-squares error over the training data set is minimized. Note that, for the case of the hyperbolic tangent activation, this minimization requires a gradient descent search, while for the linear activation case it is possible to use the Moore-Penrose pseudoinverse [1].

4 Experiments

This section presents the performance obtained from the application of the combination of RA-WE ensembles in seven binary problems: *Abalone*, *Contraceptive*, *Image*, *Kwok*, *Phoneme*, *Spam* and *Tictactoe* (the references for each data set can be found in [3]). Table 1 summarizes the main features of these problems: *dim*: number of dimensions; n_1/n_{-1}: number of samples of each class; the error rate achieved by a Support Vector Machine (SVM) with Gaussian Kernel[1]; and the recognized record error rate together with the machine that offers it (RA, RA-WE CV[2] or, when it is known, the Bayes solution).

Problem	*dim*	Train samples (n_1/n_{-1})	Test samples (n_1/n_{-1})	SVM error rate	Record error rate
Ab	8	1238/1269	843/827	20.9	19.16 (RA-WE CV)
Co	9	506/377	338/252	28.61	28.50 (RA-WE CV)
Im	18	821/1027	169/293	3.47	2.25 (RA)
Kw	2	300/200	6120/4080	11.74	11.3 (Bayes)
Ph	5	952/2291	634/1527	15.35	13.59 (RA-WE CV)
Sp	57	1673/1088	1115/725	7.2	5.69 (RA)
Ti	9	199/376	133/250	1.7	1.47 (RA)

Table 1: Characteristics of the benchmark problems.

For each problem we trained 11 different RA-WE ensembles with linearly spaced λ values between 0 and 1. The number of rounds (T) for each problem was adjusted first for $\lambda = 0.5$, stopping the building of the corresponding ensemble (conventional RA in this case) when the mean value[3] of α_t was very close to 0 ($\alpha \approx 0.01$). This value was then used for all RA-WE ensembles using other values of λ.

In all cases, we have used Multi-Layer Perceptrons (MLPs) as the base learners elements, considering two numbers of hidden units (M): one corresponds to the RA design that offers the best performance; the other is chosen as the maximum value that guarantees that there are at least 25 training samples for each free parameter. MLPs training consists on randomly initializing their weights in interval $[-1, 1]$, and then, by means of a back-propagation algorithm with a unique learning step 0.01, updating them to minimize (5).

Committee output weights training depends on the activation function that is being used. If linear, the Moore-Penrose pseudoinverse solution is applied; if a hyperbolic tangent, we follow a gradient descent algorithm with a learning step

[1]Their parameters, kernel dispersion and penalty factor, are chosen by a 5 fold cross validation process.

[2]RA-WE using cross-validation for selecting λ [5].

[3]This mean value is calculated over 50 independent runs.

fixed to 0.05 during 50 epochs and decreasing linearly from 0.05 to 0 along 50 more epochs.

In Table 2 we present the average error rates (50 runs) that are obtained by the new approach (both when using linear and non-linear activations), and compare them to those achieved by classical RA algorithm and RA-WE CV (the RA − WE algorithm when λ value is selected among the set of values $\{\lambda_0 = 0, \lambda_1 = 0.1, ..., \lambda_{10} = 1\}$ by a five-fold cross validation process [5]). Besides, the results corresponding to the RA − WE network that achieves the best performance in test, denoted as RA − WE_0, are also presented. This "omniscient" approach is not valid for designing purposes, but it is interesting to evaluate the performance of the new committees.

Furthermore, to check the statistical significance of the results, we have included the value resulting of applying the Wilcoxon Rank-sum test [6]: p_{RA}, with respect to classical RA, and p_{CV}, with respect to RA-WE CV. We remark that a value below 0.1 means that there exists a statistical difference between the two approaches, while a value close to 1 reflects the opposite situation.

	M	E_{RA}	E_{CV}	E_0	Linear activation			Tanh activation		
					E_{lin}	p_{RA}	p_{CV}	E_{th}	p_{RA}	p_{CV}
Ab	9	19.43	19.40	19.18	**19.20**	0	0	19.36	0.18	0.70
	6	19.20	19.16	19.00	**18.92***	0	0	18.98	0	0
Co	4	28.90	**28.61**	28.56	28.79	0.93	0.30	29.56	0.02	0
	3	29.20	**28.50**	28.50	28.78	0.2	0.17	29.19	0.71	0
Im	9	2.25	2.25	2.25	2.48	0	0	**2.10***	0.01	0.01
	2	2.86	2.89	2.86	2.98	0	0.08	**2.67**	0	0
Kw	9	11.68	**11.63**	11.63	11.65	0.06	0	11.69	0.49	0
	4	11.82	11.72	11.70	**11.70**	0	0.24	11.71	0	0.25
Ph	36	13.89	13.59	13.56	13.61	0.03	0.81	**13.48**	0	0.27
	28	13.70	13.60	13.52	12.33	0	0	**12.30***	0	0
Sp	5	5.69	5.73	5.69	**5.50***	0.06	0.05	5.79	0.38	0.31
	2	6.04	6.03	6.03	6.00	0.64	0.74	**5.92**	0.21	0.22
Ti	8	**1.47**	1.47	1.47	4.29	0	0	4.21	0	0
	2	8.12	**6.94**	6.94	8.63	0.1	0	7.95	0.9	0.02

Table 2: Error rates achieved by the committees of RA-WE ensembles using linear and non-linear activations (E_{lin} and E_{th}, respectively) compared to RA (E_{RA}), RA-WE CV (E_{CV}), and the "omniscient" reference RA − WE_0 (E_0).

These results show that both combination methods not only improve RA and RA-WE CV accuracy in most the problems, but even that the RA − WE_0 error rate is reduced significantly in four out of the seven data sets: **Ab**, **Im**,

Ph and **Sp**. Even more, record rates have been improved by the ensemble with linear activation for **Ab** and **Sp**, and by that employing the hyperbolic tangent for **Im** and **Ph** (these new records are marked with an asterisk in Table 2).

Examining in detail the other problems, specially **Ti** with $M = 2$, we have observed that some RA-WE networks present high error rates, degrading committees performances. This problem could be avoided using techniques to eliminate these "low quality" RA-WE networks, so getting additional significant error rates improvements, together with a reduction in the committee complexity.

5 Conclusions and future work

In this paper we have taken advantage of the diversity that RA-WE networks present when different emphasis trade-offs are used to build committees with them. It has been shown experimentally that significant error rate reductions can be achieved with these approaches with respect to RA and RA-WE CV.

Studies to evaluate other combination procedures and to simplify the complexity of the resulting committees of boosting ensembles are interesting research subjects along this promising avenue.

References

[1] C. Bishop, ed., *Neural Networks for Pattern Recognition*, Oxford University Press, 1996.

[2] Y. Freund and R. E. Schapire, Experiments with a new boosting algorithm. In *Proc. 13th International Conference on Machine Learning*, pages 148-156, Bari, Italy, 1996.

[3] V. Gómez-Verdejo, M. Ortega-Moral, J. Arenas-García and A. R. Figueiras-Vidal, Boosting by weighting boundary and erroneous samples. In *Proc. 13th European Symposium on Artificial Neural Networks*, pages 85-90, Bruges, Belgium, 2005.

[4] V. Gómez-Verdejo, J. Arenas-García, M. Ortega-Moral and A. Figueiras-Vidal, Designing RBF classifiers for weighted boosting. In *Proc. International Joint Conference on Neural Networks 2005*, pages 1057-1062, Montreal, Canada, 2005.

[5] V. Gómez-Verdejo, M. Ortega-Moral, J. Arenas-García and A. R. Figueiras-Vidal, Boosting by weighting critical and erroneous samples. To be published in Neurocomputing Journal.

[6] H. B. Mann, D. R. Whitney, On a test of whether one of two random variables is stochastically larger than the other. *The Annals of Mathematical Statistics*, 18(1):50-60, 1947.

[7] 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, Canberra, Australia, 2003.

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

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