

Committees of Adaboost ensembles with modified emphasis functions[★]

Vanessa Gómez-Verdejo, Jerónimo Arenas-García,
Aníbal R. Figueiras-Vidal

*Dep. Signal Theory and Communications, Universidad Carlos III de Madrid,
Avda. de la Universidad 30, 28911 Leganés (Madrid), SPAIN.*

Abstract

Real Adaboost ensembles with weighted emphasis (RA-we) on erroneous and critical (near the classification boundary) samples have recently been proposed, leading to improved performance when an adequate combination of these terms is selected. However, finding the optimal emphasis adjustment is not an easy task.

In this paper, we propose to make a fusion of the outputs of RA-we ensembles trained with different emphasis adjustments by means of a generalized voting scheme. The resulting committee of RA-we ensembles can retain the performance of the best RA-we component and even, occasionally, can improve it. Additionally, we present an ensemble selection strategy that removes from the committee RA-we ensembles with very poor performance.

Experimental results show that these committees frequently outperform RA and RA-we with cross validated emphasis.

Key words: Boosting, Adaboost, Neural Network ensembles, voting schemes.

[★] This work has been supported by Spanish Ministry of Education and Science grant CICYT TEC2008-02473.

1 Introduction

Boosting methods emerge in 1990s under the idea of combining weak learners to build an ensemble of improved performance. Since then, many boosting variants have been presented [2,4,11], and, among them, the Real Adaboost (RA) algorithm [12], proposed by Freund and Schapire in 1999. The RA algorithm works by iteratively training a set of learners, in such a way that each learner pays more attention to the patterns that are more difficult to classify according to the previous learners criteria. In order to do so, RA employs an emphasis function that indicates the weight that each learner should assign to each pattern during its training. However, a closer look at the RA emphasis function reveals that it can actually be decomposed into the product of two factors [6]. The first factor depends on the quadratic error of each sample, while the second is a function of the “proximity” of the sample to the classification border. In RA, these two emphasis factors are combined in a fixed manner; however, it should be expected that a more flexible scheme, adapted to the characteristics of each particular problem, can lead to improved performance.

In [6] we presented a generalization of the RA algorithm, where the tradeoff between both emphasis factors is controlled by means of an adjustable mixing parameter, λ . When compared to classic RA, this new version with weighted emphasis (RA-we) was found to achieve significant performance gains, provided that the mixing parameter was adequately selected. Although we have already explored the possibility of adjusting λ using cross-validation (CV) [6] or a dynamical selection procedure [5], this issue still remains an open problem.

Relying on the problem of selecting adequately the mixing parameter, in this paper, rather than trying to find its optimal value, we will combine the outputs of a set of RA-we ensembles, each one trained with a different value of λ . In other words, we will further use the idea of combining a set of learners to exploit the diversity among them. It will turn out that this committee of RA-

we ensembles¹ can retain the properties of the best RA-we network, thus offering a solution to problem of emphasis adjustment. Furthermore, we will see that the committee can occasionally outperform the best ensemble, an effect that can be explained in terms of reduction of the error variance, which leads to improved generalization capabilities [13].

The rest of the paper is structured as follows: In the next section, we will describe our RA-we method in general terms. In Section 3, we will present the proposed combination method to build committees of RA-we. In Section 4, we will test the performance of this method in some benchmark problems. Finally, conclusions are presented.

2 Real Adaboost with weighted emphasis

Let us consider a binary classification problem described by a set of L training patterns $\{\mathbf{x}^{(l)}\}_{l=1}^L$ and their corresponding labels $\{d^{(l)}\}_{l=1}^L$, where $\mathbf{x}^{(l)} \in \mathfrak{R}^N$ and $d^{(l)} \in \{-1, 1\}$. To solve this classification problem, RA-we sequentially trains, during several rounds, $t = 1, \dots, T$, a set of base learners. The outputs of all base learners are then combined to produce the overall output, according to

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

where $o_t(\mathbf{x})$ is the output of the t -th learner, which is constrained to range $[-1, 1]$, and α_t is its corresponding output weight. Each base learner is trained to minimize a weighted mean-square error (MSE) over the training data set:

$$C_t = \sum_{l=1}^L D_{\lambda,t}(\mathbf{x}^{(l)}) [d^{(l)} - o_t(\mathbf{x}^{(l)})]^2, \quad (2)$$

¹ To avoid ambiguity, in this paper we will indistinctly refer to the first layer components as (RA-we) ensembles, while the term committee will be exclusively used to denote the overall scheme resulting from their combination.

where the emphasis function $D_{\lambda,t}(\cdot)$ indicates the attention that each learner must assign to each training pattern. In contrast to the classical RA emphasis function, the RA-we emphasis function is built by means of a flexible combination of two emphasis terms: the first term takes large values for patterns with large quadratic error, and the second term increases for patterns that lie close to the boundary. In the first round, this emphasis function follows a flat distribution ($D_{\lambda,1}(\mathbf{x}^{(l)}) = \frac{1}{L}$, $l = 1, \dots, L$) and, afterwards, it is updated at each round according to

$$D_{\lambda,t+1}(\mathbf{x}^{(l)}) = \frac{1}{Z_t} \exp \left\{ \lambda [f_t(\mathbf{x}^{(l)}) - d^{(l)}]^2 - (1 - \lambda) f_t^2(\mathbf{x}^{(l)}) \right\}, \quad (3)$$

where Z_t is a normalization factor which ensures $\sum_{l=1}^L D_{\lambda,t+1}(\mathbf{x}^{(l)}) = 1$, and $\lambda \in [0, 1]$ is a constant to be selected so that the resulting mixed emphasis function better fits the characteristics of each particular problem. If $\lambda = 1$, RA-we only pays attention to the most erroneous patterns, while selecting $\lambda = 0$ makes the ensemble focus on critical patterns only; intermediate values of λ provide different trade offs between both kinds of emphases, with the original RA emphasis being recovered for $\lambda = 0.5$.

RA-we obtains the base learners output weights, $\{\alpha_t\}_{t=1}^T$, as:

$$\alpha_t = \frac{1}{2} \ln \left(\frac{1 + \delta_t}{1 - \delta_t} \right), \quad (4)$$

where δ_t is a generalized version of the learner's edge, which is given by the following weighted average:

$$\delta_t = \frac{\sum_{l=1}^L \exp[-f_{t-1}(\mathbf{x}^{(l)})d^{(l)}] o_t(\mathbf{x}^{(l)})d^{(l)}}{\sum_{l=1}^L \exp[-f_{t-1}(\mathbf{x}^{(l)})d^{(l)}]}. \quad (5)$$

The derivation of these expressions, as well as some theoretical insight about the RA-we method, are given in [5].

RA-we has already been shown to offer significant advantages over classic RA, provided that an adequate value for the mixing parameter is used. However, the appropriate selection remains an open problem. Unlike previous works

[5,6], in this paper we do not try to find the best λ value, but we propose to combine the outputs of several RA-we networks, each of them trained with a different value of λ , to construct a committee of improved performance.

3 Committees of RA-we networks

Although there is a large number of methods for combining the outputs of several classifiers (see [7]), we have preferred here to consider one of the most simple schemes: A generalized version of classic voting schemes.

Let us consider we have already trained J RA-we ensembles using different values of λ , $\{\lambda^{(1)}, \dots, \lambda^{(J)}\}$. Let us also denote their outputs as $f^{(j)}(\mathbf{x})$, $j = 1, \dots, J$, where output $f^{(j)}(\mathbf{x})$ corresponds to the ensemble trained with mixing parameter $\lambda^{(j)}$.

Then, the generalized voting scheme consists of two steps: First, the outputs of all RA-we ensembles for each pattern \mathbf{x} (i.e, the components of a vector $\mathbf{f}(\mathbf{x}) = [f^{(1)}(\mathbf{x}), \dots, f^{(J)}(\mathbf{x})]^T$) are sorted out according to their values (for instance, from the largest component to the smallest). In this way, a vector $\mathbf{g}(\mathbf{x}) = \text{sort}[\mathbf{f}(\mathbf{x})]$ is obtained, where the components of $\mathbf{g}(\mathbf{x})$ satisfy $g^{(1)}(\mathbf{x}) \geq g^{(2)}(\mathbf{x}) \geq \dots \geq g^{(J)}(\mathbf{x})$.

Once the outputs of the RA-we networks have been sorted out, the output of the committee is calculated as:

$$F_{\text{vot}}(\mathbf{x}) = \sum_{j=1}^J w_j g^{(j)}(\mathbf{x}), \quad (6)$$

where $\mathbf{w} = [w_1, \dots, w_J]^T$ is adjusted to minimize the training MSE. Using matrix notation, the vector of optimum weights, $\mathbf{w}^* = [w_1^*, \dots, w_J^*]^T$, can be computed as

$$\mathbf{w}^* = \mathbf{G}^\dagger \mathbf{d}, \quad (7)$$

where \mathbf{G} is a $J \times L$ matrix with elements $[\mathbf{G}]_{j,l}$ corresponding to outputs $g^{(j)}(\mathbf{x}^{(l)})$, \mathbf{G}^\dagger denotes the Moore-Penrose pseudoinverse of such a matrix, and \mathbf{d} is a column vector that contains the labels of all the training data. This scheme can be seen as a generalized version of the classic majority voting scheme, since (6) could still be applied for the classic voting scheme by just selecting \mathbf{w} to be a vector with zeros in all positions except in the middle one.

Despite the fact that this combination method takes into account the individual performance of each RA-we ensemble, removing the most inadequate RA-we ensembles can still be advantageous in terms of reducing classifier complexity and improving generalization. Since RA-we ensembles which have been trained with a very inadequate value of λ can actually achieve very poor performance, we will pay attention to the RA-we accuracy to remove the RA-we components that could negatively affect the overall committee recognition capability.

As an easy measure of accuracy, we can use the number of errors of each RA-we network on the training data set:

$$E^{(j)} = \frac{1}{2} \sum_{l=1}^L | \text{sign}[f^{(j)}(\mathbf{x}^{(l)})] - d^{(l)} |. \quad (8)$$

When the training classification error of several RA-we ensembles is zero, other error measures, such as the training MSE

$$\text{MSE}^{(j)} = \frac{1}{L} \sum_{l=1}^L [d^{(l)} - f^{(j)}(\mathbf{x}^{(l)})]^2 \quad (9)$$

can be more informative. Then, our selection criterion will be primarily based on number of errors (8); however, if the classification error of two or more ensembles is zero, we will recur to (9) as an indicator of ensemble quality.

Therefore, our selection method basically consists of retaining the ensembles whose accuracy parameters [given by (8) or (9)] are below a threshold. To select this cut-off threshold we use 5-fold cross-validation (CV) process applied

Table 1

Characteristics of the benchmark problems.

Problem	# <i>dim</i>	Train samples (n_1/n_{-1})	Test samples (n_1/n_{-1})
<i>Abalone (Ab)</i>	8	1238/1269	843/827
<i>Contraceptive (Co)</i>	9	506/377	338/252
<i>Image (Im)</i>	18	821/1027	169/293
<i>Kwok (Kw)</i>	2	300/200	6120/4080
<i>Phoneme (Ph)</i>	5	952/2291	634/1527
<i>Ripley (Ri)</i>	2	125/125	500/500
<i>Spam (Sp)</i>	57	1673/1088	1115/725
<i>Tictactoe (Ti)</i>	9	199/376	133/250

exclusively to the committee output layer.

4 Experiments

Our proposed schemes have been tested over eight binary data sets: Two synthetic problems, *Kwok* and *Ripley* (from [8] and [10], respectively); five real data sets from the UCI repository [9]: *Abalone*, *Image*, *Contraceptive*, *Spam* and *Tictactoe*; and the problem *Phoneme* (taken from [1]). The main characteristics of these data sets are summarized in Table 1, which displays the dimensionality of the input data ($\#dim$) and the number of samples belonging to each class (n_1/n_{-1}) in both the training and test sets.

The base learners of all RA-we ensembles and classical RA algorithm are Multi Layer Perceptrons (MLPs) with one hidden layer composed of M neurons, using hyperbolic tangent activations for both the output and the hidden neurons. MLP weights were trained with a back-propagation algorithm using an early stopping criterion to avoid overfitting problems. M and λ were selected for each problem using a 5-fold CV process.

The number of rounds, T , for each RA-we ensemble was selected from the evolution of weights α_t . Since these weights tend to decrease during the construction of the ensemble, at some point the new base learners do not significantly affect the ensemble performance. Therefore, as a stopping criterion we have used the relative influence of the last T_{prev} learners:

$$Q = \frac{\sum_{t'=T-T_{\text{prev}}+1}^T \alpha_{t'}}{T_{\text{prev}} \sum_{t'=1}^T \alpha'_{t'}}. \quad (10)$$

Based on this measure, the growth of the RA-we ensembles was stopped for the smallest T satisfying that (10) is below a predefined threshold Q_{stop} . T_{prev} and Q_{stop} have been experimentally fixed to 10 and 0.01, respectively.

To build the committees we have trained 11 RA-we ensembles, whose mixing parameters are taken as values from 0 to 1 with a step size of 0.1; although smaller step sizes could have been considered (with a subsequent larger number of RA-we ensembles), this setting is sufficient to exploit the diversity introduced by the mixed emphasis, as checked in [5,6]. The outputs of the ensembles have then been combined according to the method described in Section 3.

To test the performance of these committees, we are going to compare them with the following algorithms: (1) Classic RA algorithm and (2) RA-we ensembles using 5-fold CV to select the mixing parameter from set $\{0, 0.1, 0.2, \dots, 1\}$; in the following, this last method will be referred as CV RA-we.

Table 2 summarizes the results of the experiments, given in terms of the classification error rate CE of each algorithm and of classifier complexity (i.e., the number of NNs that compose each scheme). Due to the non-deterministic character of the methods under study, Table 2 shows the averages of these results (\overline{CE} and \overline{T}) over 50 runs of the algorithms and their standard deviations inside brackets. For each problem, the result of the method achieving a lower \overline{CE} is shown in boldface. We also analyze the statistical significance of the results achieved by the different algorithms by means of a T-test [3]. To

Table 2

Classification error (\overline{CE}) and classifier complexity (\overline{T}) achieved by classic RA, CV RA-we, and the proposed committees (in their basic version and with nets selection).

	Classic RA		CV RA-we		Basic committee		Committee with nets selection	
	\overline{T}	\overline{CE}	\overline{T}	\overline{CE}	\overline{T}	\overline{CE}	\overline{T}	\overline{CE}
<i>Ab</i>	31.18 (± 2.55)	19.38 (± 0.15)	33.06 (± 4.17)	19.45 (± 0.17)	310.48 (± 9.26)	19.28 (± 0.15)	153.22 (± 13.86)	19.19 (± 0.14)
<i>Co</i>	33.68 (± 5.22)	29.00 (± 1.45)	26.04 (± 6.11)	28.60 (± 1.47)	321.58 (± 17.25)	28.69 (± 1.48)	30.80 (± 13.86)	25.68 (± 3.11)
<i>Im</i>	19.60 (± 2.69)	2.46 (± 0.31)	19.60 (± 2.69)	2.46 (± 0.31)	226.48 (± 13.86)	2.67 (± 0.21)	76.32 (± 24.25)	2.34 (± 0.21)
<i>Kw</i>	29.26 (± 0.99)	11.71 (± 0.05)	29.26 (± 0.99)	11.71 (± 0.05)	336.34 (± 8.63)	11.66 (± 0.02)	336.34 (± 8.63)	11.66 (± 0.02)
<i>Ph</i>	27.74 (± 2.40)	14.04 (± 0.52)	16.56 (± 0.93)	13.49 (± 0.63)	292.22 (± 14.92)	13.73 (± 0.42)	79.80 (± 24.40)	13.50 (± 0.64)
<i>Ri</i>	28.86 (± 1.28)	9.73 (± 0.09)	38.80 (± 7.43)	9.64 (± 0.19)	348.62 (± 14.57)	9.96 (± 0.07)	164.10 (± 46.67)	9.52 (± 0.14)
<i>Sp</i>	26.18 (± 4.50)	5.94 (± 0.61)	26.18 (± 4.50)	5.94 (± 0.61)	309.84 (± 22.13)	5.55 (± 0.50)	90.68 (± 19.80)	5.65 (± 0.58)
<i>Ti</i>	4490.36 (± 987.55)	0.78 (± 0.57)	6965.76 (± 1141.34)	0.89 (± 0.60)	48350.86 (± 5259.23)	2.94 (± 1.34)	26566.84 (± 4479.66)	1.32 (± 0.85)

apply this test, we calculate the following statistic

$$t_{1,2} = \frac{CE_1 - CE_2}{\sqrt{\frac{\sigma_1^2}{N} + \frac{\sigma_2^2}{N}}},$$

where N is the number of runs and $CE_{1/2}$ and $\sigma_{1/2}$ are the average classification errors and their standard deviations for the two algorithms being compared, respectively. Then, according to Student's t distribution, we can affirm that algorithm 2 outperforms algorithm 1, with a risk level of 0.05, if $t > t_{th} = 1.66$; on the contrary, method 2 is beaten by method 1 when $t < -t_{th}$. The corresponding results are presented in Table 3.

A first analysis of the performance of the basic committees shows that this method outperforms classic RA with statistical difference in four out of the eight problems (*Ab*, *Kw*, *Ph* and *Sp*), ties in *Co* ($t_{RA,Bas} = 1.05$), and loses in the remaining three cases. When the comparison is carried out with respect to CV RA-we, both methods seem to achieve quite similar results, since committees present an significant performance improvement in *Ab*, *Kw* and *Sp*, and CV RA-we wins in *Im*, *Ph*, *Ri* and *Ti*.

Table 3

Values of the T-test statistic obtained when the classification errors of classic RA, CV RA-we and committees (both in their basic version and with network selection) are compared. The t parameter subscripts indicates which algorithms are being compared, where the standard RA is denoted as RA , CV RA-we method is referred as CV and committees in their basic version and with network selection are named as Bas and Sel , respectively.

	Basic committee		Committee with network selection		
	$t_{RA,Bas}$	$t_{CV,Bas}$	$t_{RA,Sel}$	$t_{CV,Sel}$	$t_{Bas,Sel}$
Ab	3.57	5.30	6.33	8.35	3.45
Co	1.05	-0.31	6.81	6.00	6.17
Im	-3.90	-3.97	2.33	2.27	8.07
Kw	5.90	6.57	5.90	6.57	0.00
Ph	3.16	-2.24	4.70	-0.08	2.12
Ri	-15.96	-11.17	7.89	3.60	18.34
Sp	3.44	3.50	2.46	2.44	-0.92
Ti	-10.54	-9.87	-3.61	-2.92	7.19

When the network selection method is applied (last two columns of Table 2), the advantages of the committees become clearer:

- In seven of the benchmark problems (Ab , Co , Im , Kw , Ri , Ph and Sp) committees outperform classic RA. Specially significant is the \overline{CE} reduction for problem Co , where committees achieved a very important improvement over the performance of any other method.
- With respect to CV RA-we, committees yield a superior performance in six cases (Ab , Im , Kw , Ri and Sp), achieving a similar accuracy in Ph ($t_{CV,Sel} = -0.08$) and losing only in Ti .
- Finally, when both committee schemes are compared, we can conclude that application of the network selection strategy is beneficial in terms of classification accuracy; the classification error is only increased in Sp , but without statistical significance ($t_{Bas,Sel} = -0.08$). An additional advantage of removing the worst-performing RA-we ensembles from the committee is that the computational complexity of the classifier is reduced: In some cases (Co ,

Ph, Im, Sp), it can be seen that the number of MLPs is reduced around 70% of its original value for the basic committee.

The main drawback of the proposed schemes is the computational time required for training. To build a committee, 11 RA-we ensembles were trained, thus roughly increasing the computational burden by a factor of 11 with respect to standard RA ensembles (strictly speaking, it depends on the number of learners in each RA-we ensemble). Note, however, that the computational effort for committees is similar to that of CV RA-we, since the latter also require training 11 ensembles to validate the mixed emphasis parameter. The computational time for classifying new patterns is dictated by the (average) number of learners in the committee (\bar{T}). As it can be seen in Table 2 where \bar{T} values are shown, using the network selection strategy leads to smaller committees and reduced classification times.

5 Conclusions

RA-we ensembles incorporate a flexible emphasis function that allows to trade-off the attention paid to erroneous and critical samples during the construction of the ensemble. When adequately adjusted, this weighted emphasis function can lead to improved performance. In this paper, we have proposed to build committees of RA-we ensembles as a feasible solution to the adjustment of RA-we emphasis. Experiments on a set of benchmark problems show that these committees achieve (in most the cases) significant improvement in accuracy with respect to classic RA and CV RA-we.

References

- [1] P. Alinat. Periodic Progress Report 4. Technical Thomson Report TS ASM 93/S/EGS/NC/079, ROARS Project ESPRIT II - Number 5516, 1993.
- [2] L. Breiman. Arcing classifiers. *The Annals of Statistics*, 26(3):801–849, 1998.
- [3] J. P. Marques de Sá. *Applied Statistics Using SPSS, STATISTICA, MATLAB and R*. Springer-Verlag, Berlin, 2007.
- [4] Y. Freund and R. E. Schapire. Experiments with a new boosting algorithm. In *Proc. of the 13th Intl. Conf. on Machine Learning*, pages 148–156, Bari, Italy, 1996.
- [5] V. Gómez-Verdejo, J. Arenas-García, and A. R. Figueiras-Vidal. A dynamically adjusted mixed emphasis method for building boosted ensembles. *IEEE Transactions on Neural Networks*, 1:3–17, 2008.
- [6] V. Gómez-Verdejo, M. Ortega-Moral, J. Arenas-García, and A. R. Figueiras-Vidal. Boosting by weighting critical and erroneous samples. *Neurocomputing*, 69:679–685, 2006.
- [7] L. I. Kuncheva. *Combining Pattern Classifiers*. Wiley, New York, NY, 2004.
- [8] J. T. Kwok. Moderating the output of support vector classifiers. *IEEE Transactions on Neural Networks*, 10(5):1018–1031, 1999.
- [9] D. J. Newman, S. Hettich, C. L. Blake, and C. J. Merz. UCI repository of machine learning databases. <http://www.ics.uci.edu/~mllearn/MLRepository.html>, 1998.
- [10] B. D. Ripley. Neural networks and related methods for classification (with discussion). *Journal of the Royal Statistical Society*, 56(3):409–456, 1994.
- [11] G. Rätsch and M. K. Warmuth. Efficient margin maximizing with boosting. *Journal of Machine Learning Research*, 6:2131–2152, Dec. 2005.
- [12] R. E. Schapire and Y. Singer. Improved boosting algorithms using confidence-rated predictions. *Machine Learning*, 37(3):297–336, 1999.

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