Boosting d'un pool d'apprenants faibles

Henri-Maxime Suchier, Jean-Christophe Janodet,
Christine Largeron, Marc Sebban

EURISE, Faculty of Sciences, 23 rue Paul Michelon,
University of Jean Monnet, 42023 Saint-Etienne, FRANCE,
{suchierh,janodet,sebbanma,largeron}@univ-st-etienne.fr

Abstract : Nous considérons ici des problèmes d'apprentissage où les données sont présentées à l'aide de caractéristiques fortement hétérogènes, par exemple, une base de personnes où chaque individu est décrit par son nom (une chaîne de caractères), sa photo (une image), un enregistrement de sa voix (du son) et ses mensurations (des réels). Il n'existe aucun algorithme capable d'apprendre en travaillant sur toutes ces caractéristiques à la fois (sauf en utilisant un codage et donc en perdant de l'information), mais nous disposons d'algorithmes performants et spécialisés pour traiter efficacement chaque type de caractéristiques. Dans ce travail, nous proposons une nouvelle procédure de boosting, k-BOOST, permettant à ces algorithmes de collaborer activement pendant la phase d'apprentissage, et de construire ainsi une hypothèse globale très performante. Nous étudions les propriétés théoriques de k-BOOST puis nous menons des expérimentations prouvant que notre méthode fonctionne significativement mieux que toute autre combinaison d'hypothèses qui seraient construites sans collaboration.

Mots-clés : Boosting, variables hétérogènes, résultats de convergence.

1 Introduction

Most of the research on classifier ensembles aims at combining in some way the predictions of a set of *homogeneous* classifiers, that is to say, classifiers built using a single learning algorithm from various probability distributions, as done in boosting for example (Freund & Schapire, 1996; Freund & Schapire, 1997). Another approach consists in learning heterogeneous classifiers (that is, in the form of trees, neural networks, nearest-neighbors, etc.) from a single learning distribution and combining them in an efficient final classifier, as done in *stacking* (Wolpert, 1992). However, we can remark in this latter case that the notion of *heterogeneity* only characterizes the model representation, but does not concern the data themselves. In other words, what happens when each example in the learning set is described by strongly heterogeneous features such as strings, sounds, pictures, trees, ...? In fact, in their original forms, ensemble methods become either useless, which is the case for boosting that does not consider such a situation in its framework, or insufficient, for those combining classifiers built only from the subset of features that they can deal with.

However, simple examples show that such a situation can often occur. For instance, consider a dataset that describes persons with 3 features, their first name, their height and their weight, whereas the target to predict is the gender. It is clearly insufficient to use only the first name (and "forget" the other features) to achieve this task, in particular because many first names, such as "Dana", "Taylor", "Jordan", or "Claude" are shared by men and women. But on the other hand, it would be very unfortunate not to use the first name of the person and only learn the target from the 2 numerical features. Another example could be a database of on-line marketplaces such as `www.ebay.com` where the articles are described by a picture, a textual caption and a price. The variable that one would like to predict could be the interest of a specific consumer with respect to the articles. A last example is provided by the database BIOMET (Garcia-Salicetti *et al.*, 2003) which describes a person with 5 features: his face, speech, fingerprint, hand-shape and online signature. The objective being to predict whether a given person is a forger or not, the information provided by each feature is important.

The common characteristics of both these examples is that they contain heterogeneous features, *i.e.* textual information, images, sounds as well as numerical values that cannot be efficiently handled by the same learning algorithm. Indeed, on the one hand, the state of the art that allows to learn from strings usually uses n-grams (Goodman, 2001) or other grammatical inference algorithms (de la Higuera, 2004). But these techniques cannot be adapted to learn from numerical values. On the other hand, many powerful algorithms learn from numerical features but cannot deal with strings directly.

A first solution could consist in standardizing all the features into a unique type, numerical for example, thus to use an encoding stage. Even if it seems possible to change a string or an image in a quantitative vector, the main risk of such a strategy is to lose a part of the discriminant information of the feature. To overcome this drawback, another solution could consist in building an efficient (potentially boosted) classifier for each type of features and using their predictions in a global hypothesis. This idea is the one developed in a paper by Cherkauer (Cherkauer, 1996). But the main drawback of such an approach is the lack of interaction between the classifiers during the induction process. Finally, another solution could aim at using a modified version of a special case of stacked generalization, namely cascade generalization (Gama & Brazdil, 2000). The level 0 of the cascade would be built using 1 set of attributes and the dedicated learner would be used, then the level 1 would be built by combining another set of attributes with the results of the first learner, etc. However, in this case, even if there exists a collaboration between the classifiers, it is limited to a bottom-up unilateral interaction.

In this paper, we aim not only (i) at keeping the advantages of ensemble methods, based on the premise that an ensemble is often much more accurate than its individual components (Dietterich, 1997) and (ii) at dealing with heterogeneous features, but also (iii) at generating bilateral interactions between the classifiers. To achieve this task, we focus here on the adaptation of boosting to such a context. Let us recall in Algorithm 1 the strategy of boosting and its algorithm ADABOOST. ADABOOST successively trains T times a learning algorithm WL on various probability distributions $\mathbf{w}_t$ over a learning set LS that is composed of m examples. The resulting base classifiers h_t are combined into an efficient single classifier H_T. At each new round $t+1$, the current distribution exponentially favors the weights of examples misclassified by the previous classifier h_t.

for $i = 1$ **to** m **do** $w_1(x_i) = 1/m$;
for $t = 1$ **to** T **do**
  $h_t = \text{WL}(\text{LS}, \mathbf{w}_t)$;
  $\gamma_t = \sum_{i=1}^{m} w_t(x_i) y_i h_t(x_i)$;
  $c_t = (1/2) \ln((1+\gamma_t)/(1-\gamma_t))$;
  $Z_t = \sum_{i=1}^{m} w_t(x_i) \exp(-c_t y_i h_t(x_i))$;
  for $i = 1$ **to** m **do** $w_{t+1}(x_i) = w_t(x_i) \exp(-c_t y_i h_t(x_i)) / Z_t$
return H_T such that $H_T(x) = \text{sign}\left(\sum_{t=1}^{T} c_t h_t(x)\right)$

Algorithm 1: Pseudo-code of ADABOOST.

A first boosting solution to deal with heterogeneous features would aim at selecting for each feature a relevant algorithm and in optimizing its performance by using ADABOOST; At the end of all the runs, one could combine the k resulting hypotheses in some way into a global classifier. However, this idea would not be sufficient. Indeed, from a theoretical standpoint, optimizing individual performances would not ensure an optimization of the final classifier. Moreover, by boosting each weak learner *independently* on the others, the main risk would be to encounter an overfitting phenomenon or to decrease the convergence speed of the algorithm. For instance, consider a person described by his face and his voice. Let us assume that the database contains 2 twin brothers with the same face but different voices. Boosting a weak learner WL_1 from the faces, and independently a weak learner WL_2 from the voices, would result in a useless increase of the learning time of WL_1, whereas a collaboration between both would "inform" WL_1 that the discrimination is possible thanks to WL_2 and, thus, that it is useless to try to discriminate 2 similar faces.

Therefore, we think that a better way to proceed consists in learning k classifiers in parallel *at each step* of boosting, and so in taking into account all the information provided by these k classifiers in the weight update rule. This strategy requires the construction of a new weighting scheme and the verification that it preserves the boosting theoretical properties. That is the aim of this paper.

First, we propose in Section 2 a new boosting algorithm, called k-BOOST. Then, in the core of the paper, we aim at verifying that our weighting scheme keeps the standard boosting properties: In Section 3, we tackle the problem of the convergence of the empirical error; We deal with the generalization error in Section 4. Note that we only concentrate our efforts in this paper on the particular case of $k = 2$, for which we are able to provide exact theoretical results. Finally, in Section 5, we carry out many experiments to show the relevance of our approach, before concluding.

2 The Algorithm k-BOOST

Let $\text{LS} = \{(x_1, y_1), \ldots, (x_m, y_m)\}$ be a finite set of m learning examples. Each instance x_i belongs to a domain $\mathcal{X}$ and is assigned to a boolean class $y_i \in \{-1, +1\}$. We assume that LS was generated according to some fixed but unknown distribution $\mathcal{D}$ over

$\mathcal{X} \times \{-1, +1\}$. Since we suppose that each example is described with strongly heterogeneous features, $\mathcal{X}$ is thus some cartesian product $\mathcal{X}_1 \times \mathcal{X}_2 \times \ldots \times \mathcal{X}_k$. For instance, in the first example given in Section 1, LS is a set of persons described by their first name and their weight and their height, so $\mathcal{X}_1$ is a set of strings and $\mathcal{X}_2$ is the set of real numbers covering both weight and height features. In this case, there does not exist any algorithm WL able to learn a classifier from the whole space $\mathcal{X}$ but several algorithms, each of them working on a part of $\mathcal{X}$, can cover it totally (for instance, one can use n-grams to deal with $\mathcal{X}_1$ and C4.5 to learn a decision tree from $\mathcal{X}_2$). Generalizing, let us assume that we have k algorithms, denoted WL_1, ..., WL_k, which will be used on their corresponding subset of features.

for $i = 1$ **to** m **do** $w_1(x_i) = 1/m$;
for $t = 1$ **to** T **do**
  for $j = 1$ **to** k **do** $h_{jt} = \text{WL}_j(\text{LS}, \mathbf{w}_t)$;
  define function $Z_t(u_1, \ldots, u_k) = \sum_{i=1}^{m} w_t(x_i) \exp\left(-\sum_{j=1}^{k} u_j y_i h_{jt}(x_i)\right)$;
  compute $(c_{1t}, \ldots, c_{kt}) \in \mathbb{R}^k$ **such that** $Z_t(c_{1t}, \ldots, c_{kt})$ is minimum;
  let $Z_t = Z_t(c_{1t}, \ldots, c_{kt})$;
  for $i = 1$ **to** m **do** $w_{t+1}(x_i) = w_t(x_i) \exp\left(-\sum_{j=1}^{k} c_{jt} y_i h_{jt}(x_i)\right) / Z_t$
return H_T such that $H_T(x) = \text{sign}\left(\sum_{t=1}^{T} \sum_{j=1}^{k} c_{jt} h_{jt}(x)\right)$

Algorithm 2: Pseudo-code of k-BOOST.

At each step t of our boosting algorithm, called k-BOOST (see Algorithm 2), a distribution $\mathbf{w}_t$ is defined over LS. Then, each learner WL_j uses its own *view* of the data (*i.e.*, the features it can handle) and the distribution $\mathbf{w}_t$ to produce a hypothesis h_{jt}. Then $h_{1t}, \ldots, h_{kt}$ are combined into a weighted classifier whose *global* response is used to update $\mathbf{w}_t$. Finally, the resulting hypothesis H_T is a combination of all the weighted hypotheses produced by k-BOOST. Notice that when only one learner is used ($k = 1$), our algorithm is exactly ADABOOST, so the former is an extension of the latter. Moreover, concerning computation time issues, note that k-BOOST can be run in parallel. Therefore, by using k different machines, the total amount of running time should not exceed (assuming a small communication time between processors) that required by ADABOOST on the worst algorithm among $\text{WL}_1, \ldots, \text{WL}_k$.

3 Results on the Empirical Error of 2-BOOST

The empirical error $\varepsilon(H_T, \text{LS})$ is the error of H_T computed on LS. In this section, we show that $\varepsilon(H_T, \text{LS})$ is bounded by a quantity that decreases with the number of boosting iterations. Even though some of these results can be extended to k-BOOST ($\forall k$), we focus our attention on the special case of $k = 2$ for sake of simplicity.

3.1 Conditions of the Empirical Error Minimization

Let us define the empirical error:

$$\varepsilon(H_T, \text{LS}) = (1/m) \sum_{i=1}^{m} [\![H_T(x_i) \neq y_i]\!],$$

where $[\![\pi]\!]$ is 1 if predicate π holds and 0 otherwise. Running 2-BOOST, we obtain the following result:

Lemma 1 $\varepsilon(H_T, \text{LS}) \leq \left(\prod_{t=1}^{T} Z_t\right)$, *where*

$$Z_t = \sum_{i=1}^{m} w_t(x_i) \exp\left(-c_{1t} y_i h_{1t}(x_i) - c_{2t} y_i h_{2t}(x_i)\right).$$

Proof: Let $A_i = -\sum_{t=1}^{T} \left(c_{1t} y_i h_{1t}(x_i) + c_{2t} y_i h_{2t}(x_i)\right)$. Unraveling the update rule of 2-BOOST, we get $w_{T+1}(x_i) = w_1(x_i) \exp(A_i) / \left(\prod_{t=1}^{T} Z_t\right)$, so summing $w_{T+1}(x_i)$ for all $i \in 1..m$ yields $\left(\prod_{t=1}^{T} Z_t\right) = (1/m) \sum_{i=1}^{m} \exp(A_i)$. On the other hand, $[\![H_T(x_i) \neq y_i]\!] = 1$ iff $H_T(x_i) y_i = -1$, that is to say, $A_i \geq 0$. Therefore, $\exp(A_i) \geq [\![H_T(x_i) \neq y_i]\!]$. So we deduce $\varepsilon(H_T, \text{LS}) \leq (1/m) \sum_{i=1}^{m} \exp(A_i) = \left(\prod_{t=1}^{T} Z_t\right)$, that is the statement of the Lemma. □

As a consequence of Lemma 1, the smallest $Z_1, \ldots, Z_T$ are, the smallest the empirical error is. Therefore, as for ADABOOST, 2-BOOST aims at computing the values c_{1t}, c_{2t} that minimize Z_t. Since Z_t is a convex function of (c_{1t}, c_{2t}) (see (Schapire & Singer, 1998, Appendix A) for a proof), so we need to solve:

$$\left(\frac{\partial Z_t}{\partial c_{1t}}\right) = \left(\frac{\partial Z_t}{\partial c_{2t}}\right) = 0. \tag{1}$$

To tackle this problem, we first decompose Z_t by separating the elements of the sum with respect to the positive and negative values of $y_i h_{1t}(x_i)$ and $y_i h_{2t}(x_i)$. So we define, $\forall b_1, b_2 \in \{-, +\}$, the sets $E_t(b_1 b_2) = \{i \in 1..m : (y_i h_{1t}(x_i) = b_1 1) \wedge (y_i h_{2t}(x_i) = b_2 1)\}$ and their weights $W_t(b_1 b_2) = \sum_{i \in E_t(b_1 b_2)} w_t(x_i)$. Then we get:

$$\begin{aligned}
Z_t &= W_t(++)e^{-c_{1t}-c_{2t}} + W_t(+-)e^{-c_{1t}+c_{2t}} \\
&+ W_t(-+)e^{c_{1t}-c_{2t}} + W_t(--)e^{c_{1t}+c_{2t}}, && (2)\\
(\partial Z_t/\partial c_{1t}) &= -W_t(++)e^{-c_{1t}-c_{2t}} - W_t(+-)e^{-c_{1t}+c_{2t}} \\
&+ W_t(-+)e^{c_{1t}-c_{2t}} + W_t(--)e^{c_{1t}+c_{2t}} = 0, && (3)\\
(\partial Z_t/\partial c_{2t}) &= -W_t(++)e^{-c_{1t}-c_{2t}} + W_t(+-)e^{-c_{1t}+c_{2t}} \\
&- W_t(-+)e^{c_{1t}-c_{2t}} + W_t(--)e^{c_{1t}+c_{2t}} = 0. && (4)
\end{aligned}$$

We add and subtract Eq.(3) and (4) in order to solve Eq.(1), which brings $c_{1t} + c_{2t} = \frac{1}{2} \ln\left(\frac{W_t(++)}{W_t(--)}\right)$ and $c_{1t} - c_{2t} = \frac{1}{2} \ln\left(\frac{W_t(+-)}{W_t(-+)}\right)$. Therefore, we deduce:

Proposition 1
The empirical error of 2-BOOST *is minimal when* $\forall t \in 1..T$*:*

$$c_{1t} = \frac{1}{4}\ln\left(\frac{W_t(++)W_t(+-)}{W_t(--)W_t(-+)}\right), c_{2t} = \frac{1}{4}\ln\left(\frac{W_t(++)W_t(-+)}{W_t(--)W_t(+-)}\right). \quad (5)$$

Moreover, the minimal value of Z_t *is:*

$$2\sqrt{W_t(++)W_t(--)} + 2\sqrt{W_t(+-)W_t(-+)}. \quad (6)$$

Note that Eq.(5) are meaningful only if $\forall b_1, b_2 \in \{-,+\}, W_t(b_1 b_2) \neq 0$. We assume this in the following but it may be wrong in practice. In this case, 2-BOOST will have to stop and return H_{t-1}, as ADABOOST does when $W_t(+)$ or $W_t(-)$ are null (Meir & Raetsch, 2003).

3.2 The Characteristic Parameters of 2-BOOST

It is well-known that the empirical error with ADABOOST exponentially converges towards 0 with the number of iterations T. The usual way to prove this consists in showing that each Z_t is significantly < 1 for all $t \geq 1$; This is done by introducing a characteristic parameter of ADABOOST, denoted γ_t and called the *edge* of hypothesis h_t (Meir & Raetsch, 2003). The aim of this section is to reveal the proper characteristic parameters of 2-BOOST.

Let X_1 and X_2 be 2 random variables that specify the correctness of hypotheses h_{1t} and h_{2t} respectively. X_1 takes 2 values, either +1 when h_{1t} correctly classifies an example (*i.e.*, $y_i h_{1t}(x_i) = +1$), or -1 when h_{1t} makes an error ($y_i h_{1t}(x_i) = -1$). And the same for X_2 with respect to h_{2t}. In this context, the set of weights $W_t(b_1 b_2)$ describes the joint distribution of X_1 and X_2, *i.e.*, $\forall b_1, b_2 \in \{-,+\}, W_t(b_1 b_2) = \mathbb{P}[X_1 = b_1 1 \wedge X_2 = b_2 1]$. Moreover, by Eq.(2), we get $Z_t(c_{1t}, c_{2t}) = \mathbb{E}[\exp(-c_{1t}X_1 - c_{2t}X_2)]$, so Z_t is the *Laplace transform* of the random pair (X_1, X_2). For such a transform, it is known that:

$$\frac{\partial^{p+q} Z_t}{\partial^p c_{1t} \partial^q c_{2t}}(0,0) \quad = \quad (-1)^{p+q}\mathbb{E}[X_1^p X_2^q], \forall p, q \in \mathbb{N}, \quad (7)$$

where $\mathbb{E}[X_1^p X_2^q]$ is a joint moment of X_1 and X_2. In other words, Z_t is a moment-generating function that completely and uniquely determines the distribution of (X_1, X_2). Using Eq.(2) and (7), we get for all $p, q \geq 0$:

$$\begin{array}{ll} \mathbb{E}[X_1^{2p} X_2^{2q}] = \mathbb{E}[1] = 1, & \mathbb{E}[X_1^{2p+1} X_2^{2q}] = \mathbb{E}[X_1], \\ \mathbb{E}[X_1^{2p} X_2^{2q+1}] = \mathbb{E}[X_2], & \mathbb{E}[X_1^{2p+1} X_2^{2q+1}] = \mathbb{E}[X_1 X_2]. \end{array}$$

As a consequence, Z_t can be totally described with only 3 parameters: $\mathbb{E}[X_1]$, $\mathbb{E}[X_2]$ and $\mathbb{E}[X_1 X_2]$ (plus $\mathbb{E}[1] = 1$), since every higher-order moment of (X_1, X_2) is equal to one of these values.

In terms of boosting, $\mathbb{E}[X_1]$ and $\mathbb{E}[X_2]$, that we shall now denote γ_{1t} and γ_{2t}, are the edges of the hypotheses h_{1t} and h_{2t}. They quantify the relevance of both classifiers h_{1t}

and h_{2t} with respect to the class of examples, since γ_{1t} and γ_{2t} are the expected values of the correctness of the answers of h_{1t} and h_{2t} :

$$\gamma_{1t} = \mathbb{E}[X_1] = \sum_{i=1}^{m} w_t(x_i) y_i h_{1t}(x_i) \text{ and } \gamma_{2t} = \mathbb{E}[X_2] = \sum_{i=1}^{m} w_t(x_i) y_i h_{2t}(x_i). \quad (8)$$

Concerning $\mathbb{E}[X_1 X_2]$, we use it within more understandable quantities, namely the *covariance* δ_t of X_1 and X_2 and the *correlation coefficient* ρ_t of X_1 and X_2:

$$\delta_t = \mathbb{C}\text{ov}[X_1, X_2] = \sum_{i=1}^{m} w_t(x_i) h_{1t}(x_i) h_{2t}(x_i) - \gamma_{1t}\gamma_{2t}, \quad (9)$$

$$\rho_t = \frac{\mathbb{C}\text{ov}[X_1, X_2]}{\sqrt{\mathbb{V}\text{ar}[X_1]}\sqrt{\mathbb{V}\text{ar}[X_2]}} = \frac{\delta_t}{\sqrt{1-\gamma_{1t}^2}\sqrt{1-\gamma_{2t}^2}}. \quad (10)$$

Since the classifiers h_{1t} and h_{2t} collaborate for updating $\mathbf{w}_t$, it is not surprising to find ρ_t as an important parameter of 2-BOOST: It denotes the level of independence between X_1 and X_2. Other measures of independence could be used, for instance, the interclass correlation coefficient of X_2 with respect to X_1, or the χ^2-distance between X_1 and X_2, but we checked that these measures were basically related to ρ_t, due to the fact that X_1 and X_2 take only $+1$ and -1 as values. We get:

$$\begin{cases} W_t(++) + W_t(+-) + W_t(-+) + W_t(--) = 1, \\ W_t(++) + W_t(+-) - W_t(-+) - W_t(--) = \gamma_{1t}, \\ W_t(++) - W_t(+-) + W_t(-+) - W_t(--) = \gamma_{2t}, \\ W_t(++) - W_t(+-) - W_t(-+) + W_t(--) = \delta_t + \gamma_{1t}\gamma_{2t}, \end{cases}$$

$$\iff \begin{cases} W_t(++) = (\delta_t + (1+\gamma_{1t})(1+\gamma_{2t}))/4, \\ W_t(+-) = (-\delta_t + (1+\gamma_{1t})(1-\gamma_{2t}))/4, \\ W_t(-+) = (-\delta_t + (1-\gamma_{1t})(1+\gamma_{2t}))/4, \\ W_t(--) = (\delta_t + (1-\gamma_{1t})(1-\gamma_{2t}))/4, \end{cases} \quad (11)$$

by Eq.(2) and (7). Finally, plugging Eq.(11) and (10) in Eq.(6) yields:

$$\begin{aligned} Z_t &= \frac{1}{2}\sqrt{\delta_t^2 + 2\delta_t(1+\gamma_{1t}\gamma_{2t}) + (1-\gamma_{1t}^2)(1-\gamma_{2t}^2)} \\ &+ \frac{1}{2}\sqrt{\delta_t^2 - 2\delta_t(1-\gamma_{1t}\gamma_{2t}) + (1-\gamma_{1t}^2)(1-\gamma_{2t}^2)}, \\ &\text{where } \delta_t = \rho_t\sqrt{1-\gamma_{1t}^2}\sqrt{1-\gamma_{2t}^2}. \end{aligned} \quad (12)$$

3.3 Convergence of the Empirical Error

The aim of this section is to provide a bound of Z_t, that allows to show the exponential convergence of the empirical error of 2-BOOST towards 0. We first establish a *weak learning assumption* (Kearns & Vazirani, 1994; Meir & Raetsch, 2003), WLA for short, that is to say, conditions under which both WL_1 and WL_2 are *weak learners*:

Definition 1
Let LS $= \{(x_1, y_1), \ldots, (x_m, y_m)\}$ *be a finite set of* m *learning examples. An algorithm* WL *is a weak learner with respect to* LS *iff there exists a constant* $\Gamma > 0$ *such that for all distributions* $\boldsymbol{d}$ *over* LS *and all hypotheses* $h = \text{WL}(\text{LS}, \boldsymbol{d})$*,*

$$\sum_{i=1}^{m} d(x_i) y_i h(x_i) \geq \Gamma.$$

Assuming that WL_1 and WL_2 are both weak learners implies that there exist 2 constants Γ_1, Γ_2 such that for all $t \geq 1$, $\gamma_{1t} \geq \Gamma_1 > 0$ and $\gamma_{2t} \geq \Gamma_2 > 0$.

Let us now study the conditions of convergence of the empirical error. We can afford to state a simple majoration of Z_t, considered as a function of the correlation ρ_t when γ_{1t} and γ_{2t} are fixed. Indeed, let us fix particular values for γ_{1t} and γ_{2t} and study the shape of Z_t. In Figure 1, interestingly, we can observe that Z_t is dramatically smaller than 1 whatever the value of ρ_t, which ensures convergence of $\varepsilon(H_T, \text{LS})$ towards 0. Of course, we tested many configurations of γ_{1t} and γ_{2t} and that the behavior of Z_t remained the same.

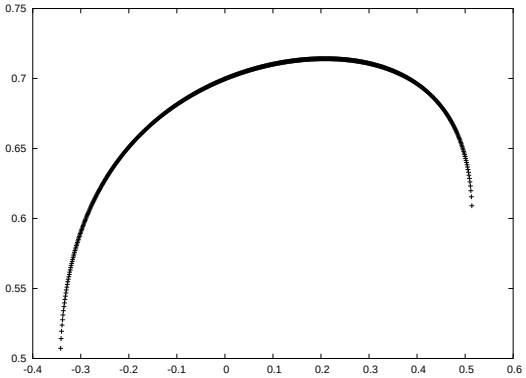


Figure 1: Z_t in function of ρ_t when $\gamma_{1t} = 0.2$ and $\gamma_{2t} = 0.7$.

Under the WLA and using Maple 9.5©, we can formally confirm the previous remarks. Indeed, starting with Eq.(12), we get: (i) when $0 < \gamma_{1t} \leq \gamma_{2t} < 1$, Z_t reaches a maximum, $\sqrt{1 - \gamma_{2t}^2}$, in $\rho_t = \frac{\gamma_{1t}}{\gamma_{2t}} \sqrt{\frac{1-\gamma_{2t}^2}{1-\gamma_{1t}^2}}$, and (ii) when $0 < \gamma_{2t} < \gamma_{1t} < 1$, Z_t reaches a maximum, $\sqrt{1 - \gamma_{1t}^2}$, in $\rho_t = \frac{\gamma_{2t}}{\gamma_{1t}} \sqrt{\frac{1-\gamma_{1t}^2}{1-\gamma_{2t}^2}}$. In other words, we get:

$$Z_t \leq \sqrt{1 - \max(\gamma_{1t}, \gamma_{2t})^2}. \tag{13}$$

Amazingly, ρ_t does not appear in this bound, *i.e.*, the empirical error of 2-BOOST is not influenced by the correlation between h_{1t} and h_{2t}, but that will not be the case for the generalization error. Let $\Gamma_0 = \max(\Gamma_1, \Gamma_2)$. We deduce:

$$Z_t \leq \sqrt{1 - \Gamma_0^2} < \exp\left(-\frac{\Gamma_0^2}{2}\right) < 1.$$

Therefore, by Lemma 1, we can conclude:

Proposition 2
Under the WLA, $\varepsilon(H_T, \text{LS}) < \exp\left(-T\frac{\Gamma_0^2}{2}\right)$, *where* $\Gamma_0 = \max(\Gamma_1, \Gamma_2)$. *So, the empirical error of 2-*BOOST *converges* $\rightarrow 0$ *when* $T \rightarrow +\infty$.

Also notice that Def.1 specifies a weak learner WL with respect to *all* distributions $\mathbf{d}$ that may be used over LS. Basically, we could only focus on the distributions $\mathbf{w}_t$. In fact, this definition allows to compare the convergence speed of ADABOOST and 2-BOOST: Let ε_{1T} (resp. ε_{2T}) be the empirical error of the classifier produced by ADABOOST when run on LS with WL_1 (resp. WL_2). It is standard to show that $\varepsilon_{1T} < \exp(-T\Gamma_1^2/2)$ and $\varepsilon_{2T} < \exp(-T\Gamma_2^2/2)$. As $\varepsilon(H_T, \text{LS}) < \exp(-T\Gamma_0^2/2)$ with $\Gamma_0 = \max(\Gamma_1, \Gamma_2)$, we conclude that:

Proposition 3
The convergence speed of 2-BOOST*, run with both* WL_1 *and* WL_2*, cannot be worse than the worst convergence speed of* ADABOOST*, run with* WL_1 *and* WL_2 *independently.*

However, in practice, we have observed that the behavior of 2-BOOST was often closer to the average of that of ADABOOST on WL_1 and WL_2 rather than the worst among them.

4 Convergence of the Generalization Error

The generalization error of the final classifier produced by ADABOOST is often observed to decrease with the number T of iterations. (Schapire *et al.*, 1998) explained this phenomenon by relating the generalization error and the margins of the learning examples. More sophisticated but realistic bounds were recently proposed in order to provide quantitative explanations (Koltchinskii & Panchenko, 2002). In this section, we recall these results and extend them to 2-BOOST.

4.1 Decomposition of the Generalization Error

Let $\mathcal{H}$ be a class of binary classifiers of VC-dimension $d_{\mathcal{H}}$. Let $\text{co}(\mathcal{H})$ denote the convex hull of $\mathcal{H}$, *i.e.*, the set of all (finite) linear combinations of hypotheses: $\text{co}(\mathcal{H}) = \{f = \sum_i \alpha_i h_i : \forall i, \alpha_i \geq 0 \text{ and } \sum_i \alpha_i = 1\}$. Given a particular $f \in \text{co}(\mathcal{H})$ and an instance x, $f(x) = \sum_i \alpha_i h_i(x)$ is a real in $[-1, +1]$; Its sign, $+1$ or -1, determines the class assigned by f to x; The *margin* $|f(x)|$ is a measure of the confidence that f gives on its prediction of the class of x.

It was proved in (Koltchinskii & Panchenko, 2002) that, given a sample $\text{LS} = \{(x_1, y_1), \ldots, (x_m, y_m)\}$ of m learning examples, drawn independently from some distribution $\mathcal{D}$ over $\mathcal{X} \times \{-1, +1\}$, and with probability at least $1 - \delta$, for all $f \in \text{co}(\mathcal{H})$ and $\theta > 0$, the *generalization error* of f is smaller than:

$$\varepsilon^{\theta}(f, \text{LS}) + \mathcal{O}\left(\frac{1}{\theta}\sqrt{\frac{d_{\mathcal{H}}}{m}}\right) + \mathcal{O}\left(\sqrt{\frac{\log(1/\delta)}{m}}\right). \tag{14}$$

The first term above, $\varepsilon^{\theta}(f, \text{LS})$, is the *empirical margin-error* of f on LS; it denotes the proportion of learning examples that are either misclassified, or correctly classified but with a small margin θ:

$$\varepsilon^{\theta}(f, \text{LS}) \quad = \quad \frac{1}{m} \sum_{i=1}^{m} [\![y_i f(x_i) \leq \theta]\!].$$

The remainder of Ineq.(14) is a complexity penalty term. The bound by (Koltchinskii & Panchenko, 2002) improves that given in (Schapire *et al.*, 1998) by removing a factor $\sqrt{\log m}$. It is rather clear that if f is able to achieve large margins on LS, then θ and δ can be chosen large, so the right-hand side of Ineq.(14), thus the generalization error of f itself, is small.

4.2 The Case of 2-BOOST

The result above holds for all voting methods, thus also for 2-BOOST[1]:

$$f_T(x) \quad = \quad \frac{\sum_{t=1}^{T} (c_{1t} h_{1t}(x) + c_{2t} h_{2t}(x))}{\sum_{t=1}^{T} (c_{1t} + c_{2t})}. \tag{15}$$

However, 2-BOOST has remarkable properties. On the one hand, it uses a special space $\mathcal{H}$ of hypotheses, that is the union of $\mathcal{H}_1$ and $\mathcal{H}_2$, the respective spaces explored by WL_1 and WL_2. From the definition of the VC-dim, we deduce that $d_{\mathcal{H}} = \min(d_{\mathcal{H}_1}, d_{\mathcal{H}_2})$. So, up to constants, the penalty term in Ineq.(14) is the same as that of the best run of ADABOOST on WL_1 and WL_2.

On the other hand, we claim that the empirical margin-error decreases with the number of iterations. Indeed, we get:

Lemma 2
$\varepsilon^{\theta}(f_T, \text{LS}) \leq \left(\prod_{t=1}^{T} Z_{\theta,t} \right)$, *where* $Z_{\theta,t} = Z_t W_t(++)^{\theta/2} W_t(--)^{-\theta/2}$.

Proof: Let $A_i = -\sum_{t=1}^{T} (c_{1t} y_i h_{1t}(x_i) + c_{2t} y_i h_{2t}(x_i))$ and $B = \theta \sum_{t=1}^{T} (c_{1t} + c_{2t})$. From Eq.(15), we deduce that $[\![y_i f_T(x_i) \leq \theta]\!] = 1$ iff $A_i + B \geq 0$, which brings $\exp(A_i + B) \geq [\![y_i f_T(x_i) \leq \theta]\!]$. Therefore, $\varepsilon^{\theta}(f_T, \text{LS}) \leq (1/m) \sum_{i=1}^{m} \exp(A_i + B)$ $= \exp(B) \left(\prod_{t=1}^{T} Z_t \right)$, by the proof of Lemma 1. Finally, since $c_{1t} + c_{2t} = (1/2)$ $\ln(W_t(++)/W_t(--))$, we get $\exp(B) = \left(\prod_{t=1}^{T} W_t(++)^{\theta/2} W_t(--)^{-\theta/2} \right)$, that allows us to conclude. □

Let us assume for the moment that the hypotheses h_{1t} and h_{2t} are independent ($\rho_t \simeq 0$). Such an assumption is often formulated in order to prove the efficiency of ensemble

[1] Since $H_T(x) = \text{sign}(f_T(x))$. Notce that Eq.(15) is sound, *i.e.*, $c_{1t} + c_{2t} > 0, \forall t \geq 1$. Indeed, $c_{1t} + c_{2t} = (1/2) \ln(W_t(++)/W_t(--))$ and $W_t(++) - W_t(--) = (\gamma_{1t} + \gamma_{2t})/2$ by Eq.(11), so $W_t(++) > W_t(--) > 0$ under the WLA.

methods (Dietterich, 2000). In such a case, by Eq.(11) and (12), we have:

$$\left\{ \begin{array}{lcl} Z_t & \simeq & \sqrt{(1-\gamma_{1t}^2)(1-\gamma_{2t}^2)}, \\ W_t(++) & \simeq & \frac{(1+\gamma_{1t})(1+\gamma_{2t})}{4}, \\ W_t(--) & \simeq & \frac{(1-\gamma_{1t})(1-\gamma_{2t})}{4}. \end{array} \right.$$

So by Lemma 2, we get:

$$Z_{\theta,t} \simeq (1+\gamma_{1t})^{\frac{1+\theta}{2}}(1-\gamma_{1t})^{\frac{1-\theta}{2}}(1+\gamma_{2t})^{\frac{1+\theta}{2}}(1-\gamma_{2t})^{\frac{1-\theta}{2}}.$$

It can be shown (Schapire *et al.*, 1998) that if $\theta < \gamma_{1t}/2$, then $(1+\gamma_{1t})^{\frac{1+\theta}{2}}(1-\gamma_{1t})^{\frac{1-\theta}{2}} < 1$ (and the same for γ_{2t}). So we conclude:

Proposition 4
*Given a fixed margin θ, if at each iteration of 2-*BOOST*, the hypotheses produced are (i) independent ($\rho_t \simeq 0$) and (ii) their respective edges γ_{1t} and γ_{2t} are $> 2\theta$, then $Z_{\theta,t} < 1$. So the empirical-margin error $\varepsilon^{\theta}(f_T, \mathrm{LS})$ of 2-*BOOST *converges towards 0 with the number of iterations (by Lemma 2).*

The generalization error of f_T will thus decrease with the number of iterations, by Ineq.(14), that will be confirmed from an experimental standpoint in Section 5.

4.3 Discussion on the Independence Assumption

By assuming the independence of the hypotheses at each round of 2-BOOST, we have shown that $Z_{\theta,t} < 1$, and we have deduced that $\varepsilon^{\theta}(f_T, \mathrm{LS})$ converged towards 0. This independence assumption could be perceived as being too strong from a practical point of view. Nevertheless, we are going to show that it could be discarded without challenging the convergence of the generalization error.

In Figure 2, we show the shape of $Z_{\theta,t}$ in function of the correlation coefficient ρ_t for fixed values of γ_{1t}, γ_{2t} and θ. Note here again that we tested several values confirming a similar behavior as the one observed in Figure 2.

From this chart, we can make the following remarks:

1. It is rather clear that when ρ_t is around 0, as we assumed in Prop. 4, $Z_{\theta,t} < 1$.

2. Moreover, we can notice that 2-BOOST will also behave well on new data if ρ_t is often strongly positive. Indeed, in such a case, h_{1t} and h_{2t} agree on the label of almost all the learning examples, so these classifiers will probably have the same behavior in the presence of new examples. However, the relevance of using 2-BOOST is limited in this case, since it has the same behavior as ADABOOST working with either WL_1 or WL_2.

3. Finally, the only case which challenges our framework occurs when ρ_t is strongly negative. Actually, in such a context, we can observe that $Z_{\theta,t} \gg 1$. This is not surprising, since $\rho_t \simeq -1$ means that the hypotheses h_{1t} and h_{2t} disagree on the class of almost all learning examples. If this often happens during the iterations of

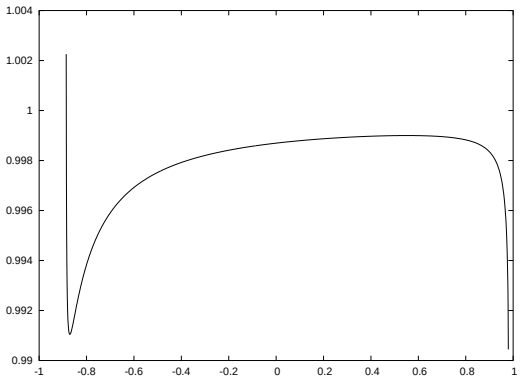


Figure 2: $Z_{\theta,t}$ in function of ρ_t when $\gamma_{1t} = 0.05$, $\gamma_{2t} = 0.07$ and $\theta = 0.02$; $Z_{\theta,t}$ becomes infinite when the correlation coefficient ρ_t becomes negative enough.

2-BOOST, then the global hypothesis f_T, that results of the combination of all h_{1t} and h_{2t}, will certainly perform randomly on any new data. However, in practice, we never faced a so strongly negative correlation between the hypotheses.

5 Experimental Results

We present in this section the experiments we carried out in order to assess the generalization abilities of 2-BOOST. In particular, we aim at showing that the global hypothesis produced by 2-BOOST from 2 learning algorithms WL_1 and WL_2 is better on average than any combination of hypotheses produced by ADABOOST from WL_1 and WL_2 independently run. To achieve this task, we will test 2 combination methods:

Method A: Both weak learners are boosted individually with ADABOOST; let $f_T(x) = (\sum_{t=1}^{T} c_t h_t(x))/(\sum_{t=1}^{T} c_t)$ and $f'_T(x) = (\sum_{t=1}^{T} c'_t h'_t(x))/(\sum_{t=1}^{T} c'_t)$ be the resulting classifiers; Method A consists in returning the sign of $f_T(x) + f'_T(x)$.

Method B: The same as Method A, except that the voting method returns the sign of the weighted combination $(\sum_{t=1}^{T} c_t) f_T(x) + (\sum_{t=1}^{T} c'_t) f'_T(x))$.

5.1 Results on a Simulated Database

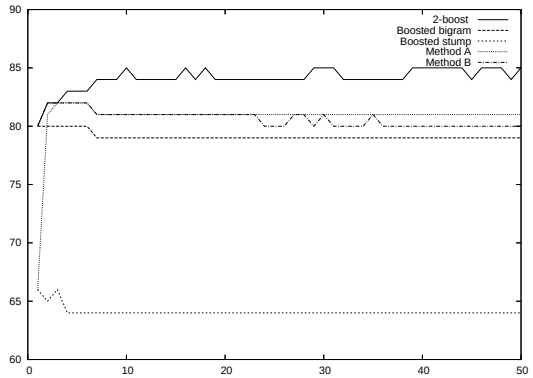
The aim of this section is to show the relevance of our approach in the presence of data described with strongly heterogeneous features. Due to the lack of unpreprocessed datasets on the web, we have decided to build a database such that any single attribute is not informative enough to learn the whole concept.

We started with a base containing 1877 first names and their associated gender: +1 for female first names and -1 for male first names. It is clearly impossible to learn the gender of a person by using only his first name, due to the overlap between both classes. Then we added a new feature, a favorite sport that could be Dance, Tennis or Soccer, by assuming that Dance was often preferred by women, Soccer by men and Tennis by

both genders. In consequence, this new feature does not allow us to deduce the gender of any person again.

Then the task consists in verifying if it is possible to build a classification model predicting the gender (+1 or -1) of a person in function of his first name and favorite sport. We consider 2 weak learners. The discrete feature (favorite sport) is investigated with a decision stump. Concerning the first names, that are strings, we designed a weak learner based on bigrams (Goodman, 2001). Roughly speaking, 2 bigrams are built, 1 per class ($+1$ and -1), that allows us to assess the probability of any string relatively to each gender. The label of any new string is then assigned by the bigram that maximizes this probability. Although the principle is rather simple, notice that we adapted this algorithm to also take into account the current distribution $\mathbf{w}_t$ and so to be a weak learner.

Figure 3 presents the results we obtained (with a 5 fold cross-validation procedure) over 50 iterations with (i) 2-BOOST, (ii) the 2 single boosted weak learners, and (iii) their combinations by Methods A and B. We can make the following remarks. First, we note that both Methods A and B outperform each single boosted algorithm, not only in terms of generalization accuracy but also of empirical accuracy, that means that each feature is useful to learn a specific part of the target concept. Moreover, 2-BOOST outperforms both Methods A and B, that proves the relevance of our boosting scheme with respect to combining independently-run algorithms. Its advantage is statistically significant using a Student paired t-test.



Algo	Empirical Accuracy	Generalization Accuracy
2-BOOST	97.73	85.10
Boosted stump	66.11	64.10
Boosted bigram	90.20	79.22
Method A	90.26	81.57
Method B	92.12	80.87

Figure 3: The curves represent the evolution over 50 iterations of the generalization accuracy using 2-BOOST, a *Boosted stump*, a *Boosted bigram*, *Method A* and *Method B*. The table shows the average results after 50 iterations of the empirical accuracy and the generalization accuracy.

5.2 A Comparison from Feature Subsets

In a second series of experiments, we verified that the behavior observed in the previous section was not an artifact due to the database. Therefore, we have used 13 databases coming from the UCI Repository[2]. Since most of them are homogeneous, we have simulated heterogeneity by randomly splitting the set of features into 2 disjoint subspaces ($\mathcal{X}_1$, $\mathcal{X}_2$) of equal size. We have run 2-BOOST with 2 weak learners: A decision stump algorithm and a naive bayesian learner (John & Langley, 1995). Table 1 (column Expe. 5.2) shows the results we get in this setting.

For each database, we present its size $|\text{LS}|$, its number of original features #Feat, and the generalization accuracy (by 5 fold cross-validation) we obtained for 2-BOOST, Method A and Method B. Moreover, we indicated in underlined font, the method which reached the best result. From this table we can make the following remarks. First, for 9 databases (over 13), our boosting procedure has the best behavior, versus 4 times for Method B and none for A. Moreover, we computed the average accuracy, by weighting each individual accuracy by the learning set size. 2-BOOST reaches a rate of 82.70%, that is much higher than 75.97% of the Method A (+6.73 in favor of 2-BOOST) and significantly higher (using a Student paired t-test) than 81.19% of the Method B (+1.51).

By analyzing the results according to the learning set size, we can also make the interesting following remark. The advantage of 2-BOOST in comparison with the Method B (which is the closest) seems to be higher on average for small databases. Actually, the average accuracy for databases containing less than 2000 instances is about 77.8% for 2-BOOST and 75.6% for the Method B (+2.2), while this difference is only of +1.3 for databases with more than 2000 instances. This result brings to the fore the necessity, above all with few examples, of a collaboration throughout the learning between both classifiers.

			Expe. 5.2			Expe. 5.3		
Base	$\|\text{LS}\|$	#Feat	2-BOOST	Method A	Method B	2-BOOST	Method A	Method B
Bigpole	1996	5	67.59	62.32	63.48	68.04	67.53	67.48
Horse	1468	23	79.90	73.50	78.68	85.35	76.63	84.60
Austral	2756	15	86.97	73.00	86.39	87.26	87.84	87.45
Balance	2496	5	92.05	71.39	89.51	98.10	97.14	97.46
Breast	2792	10	96.24	95.88	96.67	97.39	96.10	96.45
German	1004	25	73.10	73.30	73.60	73.10	73.30	73.60
Glass	167	10	74.40	72.81	72.61	81.65	79.95	81.03
Heart	274	14	79.19	79.17	79.91	81.02	81.02	78.81
Ionosphere	736	35	98.91	92.67	93.08	92.26	91.03	91.03
Pima	3068	9	73.01	72.62	72.62	73.01	72.62	72.62
TicTacToe	2396	10	78.96	71.62	74.96	91.95	90.19	92.41
WhiteHouse	439	17	96.89	95.80	95.05	98.30	97.12	97.41
xd6	604	11	74.83	70.86	75.33	75.82	75.49	75.49
Average	1728	14	82.70	75.97	81.19	85.60	84.34	85.22

Table 1: Comparison of 2-BOOST with Methods A and B on 14 databases. In Expe. 5.2, each weak learning algorithm is run from a subset of the original features. In Expe. 5.3, each weak algorithm is run with the entire set of features.

[2] http://www.ics.uci.edu/∼mlearn

5.3 Results from the Entire Feature Set

In this last series of experiments, we wanted to verify if 2-BOOST remains efficient, relatively to Methods A and B, in the case of *homogeneous* data. In other words, what happens when the whole set of features was used by both learning algorithms? Is it still relevant to use 2-BOOST? Table 1 (column Expe. 5.3) shows the results we obtained by 5 fold cross-validation.

First of all, we can note that the difference, in favor of our approach, between 2-BOOST and Methods A and B is considerably reduced. This behavior is not surprising since the 3 methods have now access to the entire database, then to more information. However, despite this, note that the difference remains statistically significant using a Student paired t-test between 2-BOOST and Methods A and B. Moreover, these results confirm the relevance and the stability of our method since 10 times over 13 it obtains the best result.

6 Discussion and Future Work

As far as we know, 2-BOOST is the first boosting procedure able to deal with heterogeneous features, so our results are encouraging. Nevertheless, a piece of work remains to be done when $k \geq 3$, particularly for proving theoretical convergence properties. A first problem is the computation of the optimal values $c_{1t}, \ldots, c_{kt}$ that minimize Z_t that can only be *approximated* by using a standard Newton-Raphson method. As for the probabilistic interpretation of Z_t as a Laplace transform, it reveals that $2^k - 1$ basic moments are required to describe Z_t. That makes things hard to prove, but we think that both the empirical and the generalization error of k-BOOST should also decrease exponentially with the number of iterations.

References

CHERKAUER K. (1996). Human expert-level performance on a scientific image analysis task by a system using combined artificial neural networks. In *Working Notes of the AAAI Workshop on Integrating Multiple Learned Models*, p. 15–21.

DE LA HIGUERA C. (2004). A bibliographic survey on grammatical inference. *Pattern Recognition*. Accepted.

DIETTERICH T. (1997). Machine learning research: for current directions. *AI Magazine*, **18**(42), 97–136.

DIETTERICH T. G. (2000). Ensemble methods in machine learning. In *First International Workshop on Multiple Classifier Systems*, p. 1–15: LNCS 1857.

FREUND Y. & SCHAPIRE R. E. (1996). Experiments with a new boosting algorithms. In *Proc. of the 13 th International Conference on Machine Learning*, p. 148–156.

FREUND Y. & SCHAPIRE R. E. (1997). A Decision-Theoretic generalization of on-line learning and an application to Boosting. *Journal of Computer and System Sciences*, **55**, 119–139.

GAMA J. & BRAZDIL P. (2000). Cascade generalization. *Machine Learning*, **41**(3), 315–343.

GARCIA-SALICETTI S., BEUMIER C., CHOLLET G., DORIZZI B., JARDINS J. L.-L., LUNTER J., NI Y. & PETROVSKA-DELACRETAZ D. (2003). Biomet: A multimodal person authentication database including face, voice, fingerprint, hand and signature modalities. In *Fourth International Conference on Audio and Video-Based Biometric Person Authentication*.

GOODMAN J. (2001). *A bit of progress in language modeling*. Rapport interne MSR-TR-2001-72, Microsoft Research Technical Report.

JOHN G. H. & LANGLEY P. (1995). Estimating continuous distributions in Bayesian classifiers. In *Eleventh Conference on Uncertainty in Artificial Intelligence*, p. 338–345.

KEARNS M. J. & VAZIRANI U. V. (1994). *An Introduction to Computational Learning Theory*. M.I.T. Press.

KOLTCHINSKII V. & PANCHENKO D. (2002). Empirical margin distributions and bounding the generalization error of combined classifiers. *Annals of Statistics*, **30**(1), 1–50.

MEIR R. & RAETSCH G. (2003). An introduction to boosting and leveraging. In *Advanced Lectures on Machine Learning*, p. 119–184: LNCS 2600.

SCHAPIRE R. E., FREUND Y., BARTLETT P. & LEE W. S. (1998). Boosting the Margin : a new explanation for the effectiveness of Voting methods. *Annals of statistics*, **26**, 1651–1686.

SCHAPIRE R. E. & SINGER Y. (1998). Improved boosting algorithms using confidence-rated predictions. In *Proc. of the 11 th International Conference on Computational Learning Theory*, p. 80–91.

WOLPERT D. H. (1992). Stacked generalization. *Neural Network*, **5**(2), 241–259.