Boosting Classifiers built from Different Subsets of Features

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

Fundamenta Informaticae XX (2009) 1–21
IOS Press

Boosting Classifiers built from Different Subsets of Features*

Jean-Christophe Janodet

Marc Sebban

Université de Lyon, F-69003, Lyon, France

Université de Saint-Etienne, F-42000, St-Etienne, France

UMR-CNRS 5516, Laboratoire Hubert Curien

18 rue du Professeur Benoit Lauras, F-42000, St-Etienne, France

{janodet,Marc.Sebban}@univ-st-etienne.fr

Henri-Maxime Suchier

Artefacto, 11 rue Meynier, F-35700 Rennes, France

hm.suchier@artefacto.fr

Abstract. We focus on the adaptation of boosting to representation spaces composed of different subsets of features. Rather than imposing a single weak learner to handle data that could come from different sources (e.g., images and texts and sounds), we suggest the decomposition of the learning task into several dependent sub-problems of boosting, treated by different weak learners, that will optimally collaborate during the weight update stage. To achieve this task, we introduce a new weighting scheme for which we provide theoretical results. Experiments are carried out and show that our method works significantly better than any combination of independent boosting procedures.

Keywords: Machine learning, boosting, heterogeneous features, subsets of features, convergence proofs.

1. Introduction

Ensemble methods aim to combine the predictions on a learning task of a set of classifiers in order to improve the accuracy that would be obtained by a single hypothesis. As mentioned in [8], an ensemble

*This work was supported in part by the IST Programme of the European Community, under the PASCAL 2 Network of Excellence, IST-2006-216886.

method will be efficient if it is able to generate some diversity in the learned hypotheses. On the one hand, this can be achieved by combining homogeneous classifiers, i.e., built using a single learning algorithm, from various probability distributions of the considered learning problem, as done in *boosting* [10, 11], *bagging* [1], or *random forests* [2]. Another possible approach consists in learning heterogeneous hypotheses (e.g., decision trees, neural networks, nearest-neighbor-based classifiers, etc.) from a single learning distribution and combining them in an efficient final classifier, as done in *stacking* [23] for instance.

Note that in this latter case, the notion of *heterogeneity* only characterizes the model nature and 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, pictures, symbolic values or trees? In fact, in their original forms, ensemble methods become either inappropriate or insufficient.

Indeed, consider a dataset that would describe persons with three features, their first name and their height and weight, whereas the target to predict would be the gender. It is clearly insufficient to use only the first name (and omit 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 unfortunate not to use the first name of the person and only learn the target from the two numerical features, since this strategy would artificially (and unfortunately) increase the Bayesian error of the problem.

Heterogeneous features often occur in real world applications. For instance, the database BIOMET [13] describes people with their faces, voices, fingerprints, hand-shapes and online signatures. If the objective is to predict whether a given person is a forger or not, then the information provided by each feature is important. Another example is provided by the databases of on-line marketplaces such as `http://www.ebay.com` where each article is described with a picture, a textual caption and a price. To design an intelligent user interface, one could be interested in predicting the interest of a specific consumer with respect to the features of the articles. Again, omitting one attribute would be problematic.

However, heterogeneous features cannot be easily handled by the same algorithm without taking some risks to lose relevant information. For instance, the state of the art that allows one to learn from strings (or trees) is often based on n-grams [14], Hidden Markov Models [9] or algorithms that are able to model long-term dependencies. In the field of Grammatical Inference [15], new techniques based on Multiplicity Automata [7] or Partially Observable Markov Models (POMM) [4] were recently proposed and today constitute indisputable standards to learn from structured data. But all these techniques cannot be adapted to learn from numerical values.

On the other hand, very powerful algorithms have been proposed to learn from those numerical features. This is the case, for instance, of the Support Vector Machines (SVM) [3]. During the past few years, many kernels have been presented in the literature allowing the use of SVM on structured data such as strings and trees. However, those kernels (e.g., spectrum kernel, mismatch kernel or subsequence kernel [6]) require the transformation of the original data into numerical feature vectors. Therefore, even if, from a technical point of view, the use of SVM on heterogeneous features is possible, we claim that such a manner to proceed leads to the loss of relevant information, such as sequentiality properties, long-term dependencies or information on the tree structure. For this reason, we aim to keep the data in their original representation space in this paper, even if this space is constituted of both structured or numerical attributes.

More precisely, our objective is to use specific algorithms on each type of features and combine them in an optimal way by an ensemble method. Note that such a strategy has already been used in machine

learning. For instance, in [5], Cherkauer proposes to learn independently an efficient classifier for each type of features and use their predictions in a global hypothesis. However, the main drawback of such an approach is the lack of interaction between the classifiers during the induction process. Another more complex solution consists in using the so-called cascade generalization [12]. Level 0 of the cascade is built using one set of attributes and a dedicated learner; then Level 1 combines another set of features with the output of the first learner, and so on ... In this case, there actually exists a collaboration between the classifiers, but it is limited due to the fact that this interaction is bottom-up, thus only unilateral.

To allow a full interaction between the classifiers, we present in this paper an adaptation of boosting to such a context of heterogeneous features. Let us recall the strategy of boosting and its well-known algorithm ADABOOST [10] (see Algorithm 1). ADABOOST consists in successively training T times a learning algorithm WL (for weak learner) on varying probability distributions $\mathbf{w}_t$ over a learning set LS 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.

Algorithm 1 Pseudo-code of ADABOOST.

Require: A weak learner WL,
a sample $\text{LS} = \{(x_1, y_1), \ldots, (x_m, y_m)\}$ where $y_i \in \{-1, +1\}$,
the maximum number T of iterations
Ensure: The (strong) combined hypothesis H_T

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

A first boosting solution to deal with heterogeneous features would consist in 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 resulting hypotheses in some way into a global classifier. However, we will experimentally show in this paper that this idea is not optimal. Indeed, boosting each weak learner independently on the others does not allow us to take in account the relationships between the features. So the main risk is to encounter an overfitting phenomenon. Moreover, from a theoretical standpoint, the optimization of individual performances does not ensure an optimization of the final classifier.

We think that a better way to proceed consists in learning classifiers in parallel at each step of boosting, and so in taking into account all the information provided by these classifiers in the weight update

rule. This strategy requires the construction of a new weighting scheme and the verification that it conserves the boosting convergence properties. Note that even if this new boosting scheme is intrinsically dedicated to deal with heterogeneous features, its potential use in a more standard framework, where features come from an unique source, is not challenged. Indeed, we claim that our new model can overcome algorithmic drawbacks by splitting high-dimensional machine learning problems into several smaller subtasks, but strongly collaborating during the boosting process.

This article is organized as follows. As mentioned before, one of our main motivations is to enable the joint use of algorithms that are known to be efficient either on structured data (strings or trees) or numerical features. Therefore, in Section 2, we consider problems represented by two types of heterogeneous features. In this context, we present a new boosting procedure, called 2-BOOST. In Sections 3 and 4, we prove that 2-BOOST is actually a boosting algorithm that leads to the decrease of both the empirical error and the generalization error. Then we carry out experiments to show the interest of our approach in Section 5; in particular, we show that our method to combine classifiers outperforms independently-boosted classifiers. Moreover, we experimentally demonstrate that 2-BOOST remains efficient on homogeneous databases. We finally conclude the paper in Section 6. As boosting more than two weak learners in parallel is an interesting issue, we have added an Appendix where we discuss the problem.

2. The Algorithm 2-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 has been generated according to some fixed but unknown distribution $\mathcal{D}$ over $\mathcal{X} \times \{-1, +1\}$.

Each example is described with strongly heterogeneous features. So we assume that $\mathcal{X}$ is some Cartesian product $\mathcal{X}_1 \times \mathcal{X}_2$. For instance, in the first example given in Section 1, LS is a set of persons described by their first name, their weight and their height, so $\mathcal{X}_1$ is a set Σ^* of strings and $\mathcal{X}_2 = \mathbb{R} \times \mathbb{R}$ covers both the weight and height features. Let us assume that we have two algorithms, denoted WL_1 and WL_2, which will be used on their corresponding subset of features. Our new boosting algorithm, called 2-BOOST, is presented in Algorithm 2.

At each step t of 2-BOOST, a distribution $\mathbf{w}_t$ is defined over LS. Then, each learner WL_j, $j = 1, 2$, uses its own view of the data (that is to say, the features it can handle) and the distribution $\mathbf{w}_t$ to produce a hypothesis h_{jt}. Then h_{1t} and h_{2t} 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 2-BOOST.

Concerning computation time issues, notice that 2-BOOST can be run in parallel. Therefore, by using two different machines, the total amount of running time should not exceed that required by ADABOOST on the worst algorithm among WL_1 and WL_2 (assuming a small communication time between processors).

3. Theoretical Results on the Empirical Error of 2-BOOST

The empirical error $\varepsilon(H_T, \text{LS})$ is the error of H_T computed on the learning sample LS, that is, the proportion of learning examples misclassified by the combined strong hypothesis. In this section, we are

Algorithm 2 Pseudo-code of 2-BOOST.

Require: Two weak learners WL_1, WL_2,
a sample $\text{LS} = \{(x_1, y_1), \ldots, (x_m, y_m)\}$,
the maximum number T of iterations
Ensure: The (strong) combined hypothesis H_T

for $i = 1$ **to** m **do**
  $w_1(x_i) \longleftarrow 1/m$
end for
for $t = 1$ **to** T **do**
  $h_{1t} \longleftarrow \text{WL}_1(\text{LS}, \mathbf{w}_t)$
  $h_{2t} \longleftarrow \text{WL}_2(\text{LS}, \mathbf{w}_t)$
  define function $Z_t(u_1, u_2) = \sum_{i=1}^{m} w_t(x_i) \exp\left(-u_1 y_i h_{1t}(x_i) - u_2 y_i h_{2t}(x_i)\right)$
  compute $c_{1t}, c_{2t} \in \mathbb{R}$ **that minimizes** $Z_t(c_{1t}, c_{2t})$
  let $Z_t = Z_t(c_{1t}, c_{2t})$
  for $i = 1$ **to** m **do**
    $w_{t+1}(x_i) \longleftarrow w_t(x_i) \exp\left(-c_{1t} y_i h_{1t}(x_i) - c_{2t} y_i h_{2t}(x_i)\right) / Z_t$
  end for
end for
return H_T **with** $H_T(x) = \text{sign}\left(\sum_{t=1}^{T} \sum_{j=1}^{2} c_{jt} h_{jt}(x)\right)$

going to show that $\varepsilon(H_T, \text{LS})$ can be bounded by a quantity that decreases with the number of boosting iterations.

3.1. Conditions of the Empirical Error Minimization

Let us define

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

where $[\![\pi]\!]$ is 1 if predicate π holds and 0 otherwise.

Running 2-BOOST, we obtain the following result:

Lemma 3.1. $\varepsilon(H_T, \text{LS}) \leq \left(\prod_{t=1}^{T} Z_t\right)$, where

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

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 $w_{T+1}(x_i) = w_1(x_i) \exp(A_i) / \left(\prod_{t=1}^{T} Z_t\right)$. $\mathbf{w}_{T+1}$ is a distribution over LS and $w_1(x_i) = (1/m)$, so summing $w_{T+1}(x_i)$ for all $1 \leq i \leq m$ 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 that $\varepsilon(H_T, \text{LS}) \leq (1/m) \sum_{i=1}^{m} \exp(A_i) = \left(\prod_{t=1}^{T} Z_t\right)$. □

As a consequence of Lemma 3.1, the smaller $Z_1, \ldots, Z_T$, the smaller the empirical error. Therefore, as for ADABOOST, 2-BOOST aims to compute, at each round, the values of c_{1t} and c_{2t} that minimize Z_t. To solve this problem, we first establish a technical result:

Lemma 3.2. Z_t is a convex function.

Proof:
The convexity of function Z_t can be established by showing that its Hessian matrix is positive semi-definite (see [21, Appendix A]). Below, we provide a direct proof by using the definition of a convex function. Let $\mathbf{u} = (u_1, u_2)$, $\mathbf{v} = (v_1, v_2) \in \mathbb{R}^2$ and $0 \leq \theta \leq 1$.

$$\begin{aligned} Z_t\left(\theta\mathbf{u} + (1-\theta)\mathbf{v}\right) &= Z_t\left(\theta u_1 + (1-\theta)v_1, \theta u_2 + (1-\theta)v_2\right) \\ &= \sum_{i=1}^{m} w_t(x_i) \exp(-\sum_{j=1}^{2} \left(\theta u_j + (1-\theta)v_j\right) y_i h_{jt}(x_i)) \\ &= \sum_{i=1}^{m} w_t(x_i) \exp\left(\theta U(x) + (1-\theta)V(x)\right) \qquad (2) \\ &\quad \text{with } \begin{cases} U(x) = -\sum_{j=1}^{2} u_j y_i h_{jt}(x_i) \\ V(x) = -\sum_{j=1}^{2} v_j y_i h_{jt}(x_i) \end{cases} \end{aligned}$$

Since exp is a convex function, we have

$$\exp\left(\theta U(x) + (1-\theta)V(x)\right) \leq \theta \exp(U(x)) + (1-\theta)\exp(V(x)).$$

Combining this inequality with Equation (2) yields

$$Z_t\left(\theta\mathbf{u} + (1-\theta)\mathbf{v}\right) \leq \theta Z_t(\mathbf{u}) + (1-\theta)Z_t(\mathbf{v}),$$

that is the statement of the Lemma. □

Therefore, by Lemma 3.1, reducing the empirical error consists in minimizing Z_t, and thanks to Lemma 3.2, the minimization consists in finding c_{1t} and c_{2t} that zero the two first-order derivatives of Z_t:

$$\left(\frac{\partial Z_t}{\partial c_{1t}}\right) = \left(\frac{\partial Z_t}{\partial c_{2t}}\right) = 0. \qquad (3)$$

Let us investigate 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 the sets:

$$\begin{aligned} E_t(++) &= \{1 \leq i \leq m : (y_i h_{1t}(x_i) = +1) \wedge (y_i h_{2t}(x_i) = +1)\}, \\ E_t(+-) &= \{1 \leq i \leq m : (y_i h_{1t}(x_i) = +1) \wedge (y_i h_{2t}(x_i) = -1)\}, \\ E_t(-+) &= \{1 \leq i \leq m : (y_i h_{1t}(x_i) = -1) \wedge (y_i h_{2t}(x_i) = +1)\}, \\ E_t(--) &= \{1 \leq i \leq m : (y_i h_{1t}(x_i) = -1) \wedge (y_i h_{2t}(x_i) = -1)\}. \end{aligned}$$

For instance, $E_t(++)$ denotes the set of examples (x_i, y_i) which are correctly classified by both h_{1t} and h_{2t}, whereas $E_t(+-)$ is the set of examples correctly classified by h_{1t} and misclassified by h_{2t}. We also introduce the corresponding weights:

$$W_t(++) = \sum_{i \in E_t(++)} w_t(x_i),$$

and weights $W_t(+-)$ and $W_t(-+)$ and $W_t(--)$ similarly.

These weights allow us to rewrite Equation (1) and compute the first order derivatives of Z_t with respect to c_{1t} and c_{2t}:

$$\begin{aligned}
Z_t(c_{1t}, c_{2t}) &= W_t(++)e^{-c_{1t}-c_{2t}} + W_t(+-)e^{-c_{1t}+c_{2t}} \\
&+ W_t(-+)e^{c_{1t}-c_{2t}} + W_t(--)e^{c_{1t}+c_{2t}}, && (4)\\
(\partial Z_t/\partial c_{1t}) &= -W_t(++)e^{-c_{1t}-c_{2t}} - W_t(+-)e^{-c_{1t}+c_{2t}} \\
&+ W_t(-+)e^{c_{1t}-c_{2t}} + W_t(--)e^{c_{1t}+c_{2t}} = 0, && (5)\\
(\partial Z_t/\partial c_{2t}) &= -W_t(++)e^{-c_{1t}-c_{2t}} + W_t(+-)e^{-c_{1t}+c_{2t}} \\
&- W_t(-+)e^{c_{1t}-c_{2t}} + W_t(--)e^{c_{1t}+c_{2t}} = 0. && (6)
\end{aligned}$$

In order to solve Equation (3), we add and substract Equations (5) and (6), that yield:

$$\begin{aligned}
c_{1t} + c_{2t} &= \frac{1}{2}\ln\left(\frac{W_t(++)}{W_t(--)}\right), && (7)\\
c_{1t} - c_{2t} &= \frac{1}{2}\ln\left(\frac{W_t(+-)}{W_t(-+)}\right). && (8)
\end{aligned}$$

So we finally deduce the following result:

Theorem 3.1. The empirical error of 2-BOOST is minimal when for all $1 \leq t \leq T$:

$$\begin{aligned}
c_{1t} &= \frac{1}{4}\ln\left(\frac{W_t(++)W_t(+-)}{W_t(--)W_t(-+)}\right), && (9)\\
c_{2t} &= \frac{1}{4}\ln\left(\frac{W_t(++)W_t(-+)}{W_t(--)W_t(+-)}\right). && (10)
\end{aligned}$$

Moreover, the minimal value of Z_t is:

$$2\sqrt{W_t(++)W_t(--)} + 2\sqrt{W_t(+-)W_t(-+)}. \quad (11)$$

Note that Equations (9) and (10) are meaningful only if $W_t(++) \neq 0$ and $W_t(+-) \neq 0$ and $W_t(-+) \neq 0$ and $W_t(--) \neq 0$. We assume these relations in the following but they may not hold in practice. In this case, 2-BOOST will have to stop and return H_{t-1}, as ADABOOST does when $W_t(+) = 0$ or $W_t(-) = 0$, that is, when the current hypothesis h_t produced by the learner perfectly classifies (or miclassifies) the learning examples [19].

3.2. The Characteristic Parameters of 2-BOOST

It is well-known that the empirical error of ADABOOST exponentially converges towards 0 with the number of iterations T [20]. The usual way to prove it consists in showing that each Z_t is significantly $<$ 1 for all $t \geq 1$. In this case, the product of Z_t's gets closer and closer to 0, at each round of ADABOOST, thus the empirical error gets closer and closer to 0 too, by Lemma 3.1.

Showing that $Z_t < 1$ is usually done by introducing a characteristic parameter of ADABOOST, denoted γ_t and called the *edge* of the hypothesis h_t [19]. Parameter γ_t plays a central role in the *weak learning assumption* [16] that is used to prove the convergence of ADABOOST. Note that in Algorithm 1, we gave the pseudo-code of ADABOOST using parameter γ_t, rather than the historical parameter called ϵ_t [11]; both are of course related (that is, $\gamma_t = 1 - 2\epsilon_t$).

The aim of this section is to display the proper characteristic parameters of 2-BOOST. Let X_1 and X_2 be two random variables that specify the correctness of hypotheses h_{1t} and h_{2t} respectively. X_1 takes two values, either $+1$ when h_{1t} correctly classifies an example (that is, $y_i h_{1t}(x_i) = +1$), or -1 when h_{1t} makes an error (that is, $y_i h_{1t}(x_i) = -1$). Similarly, X_2 takes either $+1$ when h_{2t} correctly classifies an example, or -1 when h_{2t} makes an error.

In this context, the sets of weights W_t describe the joint distribution of X_1 and X_2:

$$\begin{aligned} W_t(++) &= \mathbb{P}[X_1 = +1 \wedge X_2 = +1] \\ W_t(+-) &= \mathbb{P}[X_1 = +1 \wedge X_2 = -1] \\ W_t(-+) &= \mathbb{P}[X_1 = -1 \wedge X_2 = +1] \\ W_t(--) &= \mathbb{P}[X_1 = -1 \wedge X_2 = -1]. \end{aligned}$$

Now let us focus on Z_t. By Equation (4), we get:

$$Z_t(c_{1t}, c_{2t}) = \mathbb{E}[e^{-c_{1t}X_1 - c_{2t}X_2}], \tag{12}$$

so Z_t is the *Laplace transform* of the random pair (X_1, X_2). Developing Z_t in power series yields:

$$Z_t(c_{1t}, c_{2t}) = \sum_{p,q \in \mathbb{N}} \frac{\partial^{p+q} Z_t}{\partial^p c_{1t} \partial^q c_{2t}}(0,0) \frac{c_{1t}^p c_{2t}^q}{(p+q)!}$$

and for such a transform, it is known that for all $p, q \in \mathbb{N}$,

$$\frac{\partial^{p+q} Z_t}{\partial c_{1t}^p \partial c_{2t}^q}(0,0) = (-1)^{p+q} \mathbb{E}[X_1^p X_2^q], \tag{13}$$

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 determines completely and uniquely the distribution of (X_1, X_2). Let us use Equation (4) to compute the different derivatives of Z_t in $(0,0)$ and plug the results into Equation (13). We get, for all $p, q \geq 0$:

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

In consequence, Z_t can be totally described with only three parameters: $\mathbb{E}[X_1]$, $\mathbb{E}[X_2]$ and $\mathbb{E}[X_1X_2]$ (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. Indeed, γ_{1t} and γ_{2t} are the expected values of the correctness of the answers of h_{1t} and h_{2t}, thus real numbers in $[-1, +1]$ that measure the difference between the proportions of correctly classified and misclassified examples:

$$\gamma_{1t} = \mathbb{E}[X_1] = \sum_{i=1}^{m} w_t(x_i) y_i h_{1t}(x_i), \tag{14}$$

$$\gamma_{2t} = \mathbb{E}[X_2] = \sum_{i=1}^{m} w_t(x_i) y_i h_{2t}(x_i). \tag{15}$$

Concerning $\mathbb{E}[X_1X_2]$, we transform it into more natural quantities: the *covariance* δ_t of X_1 and X_2 and the *correlation coefficient* ρ_t of X_1 and X_2:

$$\begin{aligned}
\delta_t &= \mathbb{C}\text{ov}[X_1, X_2] \\
&= \mathbb{E}[X_1X_2] - \mathbb{E}[X_1]\mathbb{E}[X_2] \\
&= \sum_{i=1}^{m} w_t(x_i) h_{1t}(x_i) h_{2t}(x_i) - \gamma_{1t}\gamma_{2t}, \qquad (16) \\
\rho_t &= \frac{\mathbb{C}\text{ov}[X_1, X_2]}{\sqrt{\mathbb{V}\text{ar}[X_1]}\sqrt{\mathbb{V}\text{ar}[X_2]}} \\
&= \frac{\delta_t}{\sqrt{1-\gamma_{1t}^2}\sqrt{1-\gamma_{2t}^2}}. \qquad (17)
\end{aligned}$$

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 these measures are basically related to ρ_t, due to the fact that X_1 and X_2 take only $+1$ and -1 as values.

Hence, Z_t is totally determined by γ_{1t}, γ_{2t} and δ_t (or equivalently ρ_t). So let us rewrite the minimal value of Z_t, given by Equation (11), in function of these parameters. Equations (4) and (13) yields:

$$\begin{aligned}
&\begin{cases}
W_t(++) + W_t(+-) + W_t(-+) + W_t(--) &= 1, \\
W_t(++) + W_t(+-) - W_t(-+) - W_t(--) &= \gamma_{1t}, \\
W_t(++) - W_t(+-) + W_t(-+) - W_t(--) &= \gamma_{2t}, \\
W_t(++) - W_t(+-) - W_t(-+) + W_t(--) &= \delta_t + \gamma_{1t}\gamma_{2t},
\end{cases} \\
\iff &\begin{cases}
W_t(++) &= (\delta_t + (1+\gamma_{1t})(1+\gamma_{2t}))/4, \\
W_t(+-) &= (-\delta_t + (1+\gamma_{1t})(1-\gamma_{2t}))/4, \\
W_t(-+) &= (-\delta_t + (1-\gamma_{1t})(1+\gamma_{2t}))/4, \\
W_t(--) &= (\delta_t + (1-\gamma_{1t})(1-\gamma_{2t}))/4,
\end{cases}
\end{aligned} \tag{18}$$

Finally, plugging Equations (18) and (17) in Equation (11) yields:

$$\begin{aligned} Z_t &= \frac{1}{2}\sqrt{\delta_t^2 + 2\delta_t(1+\gamma_{1t}\gamma_{2t}) + (1-\gamma_{1t}^2)(1-\gamma_{2t}^2)} \\ &+ \frac{1}{2}\sqrt{\delta_t^2 - 2\delta_t(1-\gamma_{1t}\gamma_{2t}) + (1-\gamma_{1t}^2)(1-\gamma_{2t}^2)}, \\ &\text{where } \delta_t = \rho_t\sqrt{1-\gamma_{1t}^2}\sqrt{1-\gamma_{2t}^2}. \end{aligned} \tag{19}$$

3.3. Convergence of the Empirical Error

The aim of this section is to provide a bound of Z_t, that allows us to show the exponential convergence of the empirical error of 2-BOOST towards 0. We first establish a *weak learning assumption* [16, 19], that is to say, conditions under which both WL_1 and WL_2 are *weak learners*:

Definition 3.1. Let $\text{LS} = \{(x_1, y_1), \ldots, (x_m, y_m)\}$ 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 $\mathbf{d}$ over LS and all hypotheses $h = \text{WL}(\text{LS}, \mathbf{d})$,

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

Assuming that WL_1 and WL_2 are both weak learners implies that there exist two 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. To achieve this goal, we use Equation (19) and study Z_t as a function of ρ_t assuming that γ_{1t} and γ_{2t} are constants. Omitting the technicalities, we can show that:

1. 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
2. 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:

$$Z_t \leq \sqrt{1 - \max(\gamma_{1t}, \gamma_{2t})^2}. \tag{20}$$

Note that ρ_t does not appear in this bound: The empirical error of 2-BOOST is not influenced by the correlation between h_{1t} and h_{2t} (that will not be the case of the generalization error).

We now assume that WL_1 and WL_2 are both weak learners. Therefore, there exist two 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 $\Gamma_0 = \max(\Gamma_1, \Gamma_2)$. We deduce that:

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

Therefore, by Lemma 3.1, we can conclude that:

Theorem 3.2. Under the weak learning assumption, $\varepsilon(H_T, \text{LS}) < \exp\left(-T\Gamma_0^2/2\right)$. So, the empirical error of 2-BOOST converges to 0 when $T \to +\infty$.

Note that Definition 3.1 specifies a weak learner WL with respect to *all* the distributions **d** that may be defined over LS. Basically, one should only be interested in the distributions $\mathbf{w}_t$. In fact, this definition allows us to compare the convergence speed of ADABOOST and 2-BOOST. Indeed, 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 easy 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:

Theorem 3.3. 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.

4. Convergence of the Generalization Error

The generalization error of any learnt classifier f is the probability that f misclassifies any new example. Concerning ADABOOST, one often observes that the generalization error of the final classifier decreases with the number T of iterations. In [20], the authors explained this phenomenon by relating the generalization error and the margins of the learning examples. More sophisticated but realistic bounds were proposed in order to provide quantitative explanations [18]. In this section, we recall these results and extend them to 2-BOOST.

4.1. Decomposition of the Generalization Error

Let $\mathcal{H} = \{h_1, h_2, \ldots\}$ be a class of binary classifiers of VC-dimension $d_\mathcal{H}$. Let $\text{co}(\mathcal{H})$ denote the convex hull of $\mathcal{H}$, that is, the set of all finite convex combinations of hypotheses:

$$\text{co}(\mathcal{H}) = \{f = \sum_i \alpha_i h_i : \alpha_i \geq 0 \text{ and } \sum_i \alpha_i = 1\}.$$

Notice that 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 number 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 [18] 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, that is, $\mathbb{P}_\mathcal{D}[f(x) \neq y]$, is smaller than:

$$\varepsilon^\theta(f, \text{LS}) + \mathcal{O}\left(\frac{1}{\theta}\sqrt{\frac{d_\mathcal{H}}{m}}\right) + \mathcal{O}\left(\sqrt{\frac{\log(1/\delta)}{m}}\right). \tag{21}$$

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 Expression (21) is a complexity penalty term. The bound presented in [18] improves that given in [20] 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 that Expression (21), thus the generalization error of f itself, is small.

4.2. The Case of 2-BOOST

The previous result holds for all voting methods, thus also for 2-BOOST. Indeed, the global hypothesis returned by 2-BOOST is $H_T(x) = \text{sign}(f_T(x))$ with

$$f_T(x) \quad = \quad \frac{\sum_{t=1}^{T} (c_{1t} h_{1t}(x) + c_{2t} h_{2t}(x))}{\sum_{t=1}^{T} (c_{1t} + c_{2t})}, \tag{22}$$

thus $H_T = \text{sign}(f_T)$ for some $f_T \in \text{co}(\mathcal{H})$.

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 from whom WL_1 and WL_2 select their hypotheses. By the definition of the VC-dimension [22], we deduce that $d_{\mathcal{H}} = \min(d_{\mathcal{H}_1}, d_{\mathcal{H}_2})$. So, up to constants, the penalty term in Expression (21) 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 4.1. $\varepsilon^{\theta}(f_T, \text{LS}) \leq \left(\prod_{t=1}^{T} Z_{\theta,t} \right)$, 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 Equation (22), we deduce that $[\![y_i f_T(x_i) \leq \theta]\!] = 1$ if and only if $A_i + B \geq 0$, that brings $\exp(A_i + B) \geq [\![y_i f_T(x_i) \leq \theta]\!]$. Therefore, $\varepsilon^{\theta}(f_T, \text{LS}) \leq (1/m) \sum_{i=1}^{m} \exp(A_i) \exp(B) = \exp(B) \left(\prod_{t=1}^{T} Z_t \right)$, by the proof of Lemma 3.1. Finally, since $c_{1t} + c_{2t} = (1/2) \ln(W_t(++)/W_t(--))$, we deduce that $\exp(B) = \left(\prod_{t=1}^{T} W_t(++)^{\theta/2} W_t(--)^{-\theta/2} \right)$, that yields the result. □

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 methods [8]. In such a case, by Equations (18) and (19), we have:

$$\begin{cases} Z_t & \simeq & \sqrt{(1-\gamma_{1t}^2)(1-\gamma_{2t}^2)}, \\ W_t(++) & \simeq & (1+\gamma_{1t})(1+\gamma_{2t})/4, \\ W_t(--) & \simeq & (1-\gamma_{1t})(1-\gamma_{2t})/4. \end{cases}$$

So by Lemma 4.1, 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 [20] 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 using Lemme 4.1, we conclude:

Theorem 4.1. Given a fixed margin $\theta > 0$, if at each iteration of 2-BOOST, the hypotheses produced are (1) independent ($\rho_t \simeq 0$) and (2) 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.

The generalization error of f_T will thus decrease with the number of iterations, by Expression (21), 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. In this section, we justify that it can be discarded without challenging the convergence of the generalization error.

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

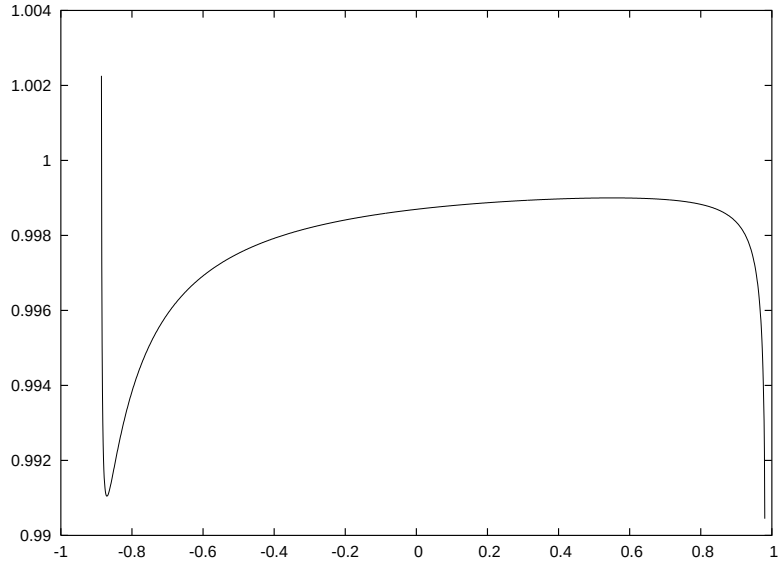


Figure 1. $Z_{\theta,t}$ as a function of ρ_t when $\gamma_{1t} = 0.05$, $\gamma_{2t} = 0.07$ and $\theta = 0.02$; notice that $Z_{\theta,t}$ becomes infinite when the correlation coefficient ρ_t is strongly negative.

We can make the following remarks. Firstly, it is rather clear that when ρ_t is around 0, as we assumed in Theorem 4.1, $Z_{\theta,t}$ is smaller than 1. 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 interest 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.

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 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, notice that 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 to show that the global hypothesis produced by 2-BOOST from two 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 two combination methods:

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

Note that, of course, many other combinaisons of classifiers could be studied, methods A and B being the most natural.

5.1. Results on the STUDENTS Database

The aim of this section is to show the relevance of our approach in the presence of data described with strongly heterogeneous features. To achieve this task, we run 2-BOOST on the database STUDENTS, that contains the marks obtained by 1877 students during sport events. Each instance is described by:

- a **string** that is the first name of the student,
- a **nominal** attribute that encodes the selected sport (Dance, Tennis or Soccer) by the student,
- an **ordinal** feature that represents the obtained mark and
- a **boolean** value that encodes the gender of the individual ($+1$ for females, -1 for males).

The learning task consists in building a classification model predicting the gender of a person in function of his first name, selected sport and mark. Some of these features seem to be partially discriminative to learn the target concept. Indeed, it is well-known that Soccer is often chosen by boys while Dance is usually selected by girls. However, Tennis can be equally chosen by both genders. On the other hand, the boys are often more interested in the practice of sports, and we can wonder if there is a statistical dependence with the obtained mark. Finally, the first names clearly give a lot of information about the gender of the individuals. However, this is insufficient to perfectly discriminate the two classes, due to

the overlap of the two considered distributions, as we explained in the introduction of this paper. So, this database is clearly interesting to test the ability of 2-BOOST to deal with heterogeneous features.

We consider two weak learners in this experimental study. The discrete features (selected sport and mark) are tackled with a decision stump. Concerning the first names, we used a bigram-based learner [14]. Roughly speaking, two bigrams are built, one 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.

Figure 2 presents the results we obtained (with a 5 fold cross-validation procedure [17]) over 50 iterations with (i) 2-BOOST, (ii) the two 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 type of features is useful to learn a subpart of the target concept. Moreover, 2-BOOST outperforms both Methods A and B, that proves the interest of our boosting scheme with respect to combining independently-run algorithms. Its advantage is statistically significant using a Student paired t-test.

5.2. Results on UCI Repository Databases

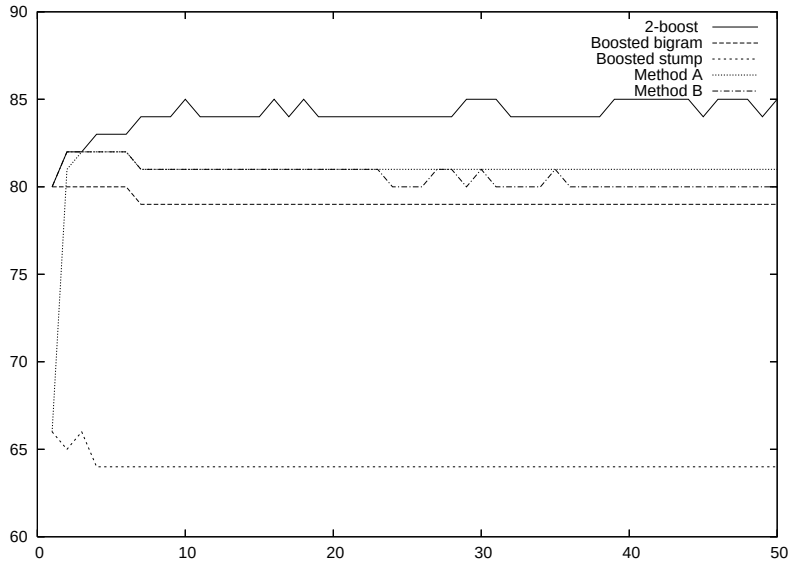
In a second series of experiments, we verified that the behavior observed in the previous section was not an artefact due to the specificity of the database. Therefore, we used 13 databases coming from the UCI Repository[1]. Since most of them are homogeneous (i.e., composed of features of the same type), we have simulated heterogeneity by randomly splitting the set of features into two 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.

Table 1 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 indicate 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 have 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 Method A ($+6.73$ pourcentage points in favor of 2-BOOST) and significantly higher (using a Student paired t-test) than 81.19% of Method B ($+1.51$ points).

By analyzing the results according to the learning set size, we can also remark that the advantage of 2-BOOST in comparison with 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 Method B ($+2.20$ points), while this difference is only of $+1.30$ points for databases with more than 2000 instances. This result brings to the fore the necessity, particularly on small datasets, of a collaboration between both classifiers.

[1] http://www.ics.uci.edu/∼mlearn



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 2. 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.

5.3. Behavior of 2-BOOST on Homogeneous Databases

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 is used by both learning algorithms? Is it still relevant to use 2-BOOST?

Table 2 shows the results we obtained by 5 fold cross-validation, using the same format as that of Table 1. 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 three methods have now access to the entire database, thus to more information. The advantage of collaborating during the learning is reduced. 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. Finally, as we did before, we computed the average accuracy according to the size of the databases. The previously mentionned behavior remains the same. Actually, despite the fact

Table 1. Comparison of 2-BOOST with Methods A and B on 13 databases. Each weak learning algorithm is run from a subset of the original features.

Base	$\lvert$LS$\rvert$	#Feat	2-BOOST	Method A	Method B
Austral	2756	15	<u>86.97</u>	73.00	86.39
Balance	2496	5	<u>92.05</u>	71.39	89.51
Bigpole	1996	5	<u>67.59</u>	62.32	63.48
Breast	2792	10	96.24	95.88	<u>96.67</u>
German	1004	25	73.10	73.30	<u>73.60</u>
Glass	167	10	<u>74.40</u>	72.81	72.61
Heart	274	14	79.19	79.17	<u>79.91</u>
Horse	1468	23	<u>79.90</u>	73.50	78.68
Ionosphere	736	35	<u>98.91</u>	92.67	93.08
Pima	3068	9	<u>73.01</u>	72.62	72.62
TicTacToe	2396	10	<u>78.96</u>	71.62	74.96
WhiteHouse	439	17	<u>96.89</u>	95.80	95.05
xd6	604	11	74.83	70.86	<u>75.33</u>
Average	1728	14	82.70	75.97	81.19

that the differences are slightly reduced, the average of 2-BOOST is higher ($+0.59$ points) for datasets containing less than 2000 instances, while its advantage is only of $+0.28$ points when there are more than 2000 examples.

6. Conclusion

As far as we know, 2-BOOST is the first boosting procedure able to deal with heterogeneous features. We provided exact theoretical results in the case of 2-BOOST and the experiments confirmed that it allows dramatic improvements in terms of accuracy with respect to any basic combinaison of the two learned classifiers.

Even if we think that 2-BOOST is sufficient to tackle a large range of machine learning problems, the case of $k > 2$ weak learners remain to be studied. In Appendix (see below), we show that the convergence proofs require the call of complex approximation methods to assess the confidence parameters used in final linear combination of the hypotheses.

Why so many efforts to prove the convergence of k-BOOST? In fact, while several numerical vectors can be actually concatenated into a single vector, the picture is less clear as soon as one considers several strings and trees. Hence, k-BOOST could be able to approach any problem with heterogeneous features.

Table 2. Comparison of 2-BOOST with Methods A and B on 13 databases. Each weak algorithm is run with the entire set of features.

Base	$\lvert$LS$\rvert$	#Feat	2-BOOST	Method A	Method B
Austral	2756	15	87.26	87.84	87.45
Balance	2496	5	98.10	97.14	97.46
Bigpole	1996	5	68.04	67.53	67.48
Breast	2792	10	97.39	96.10	96.45
German	1004	25	73.10	73.30	73.60
Glass	167	10	81.65	79.95	81.03
Heart	274	14	81.02	81.02	78.81
Horse	1468	23	85.35	76.63	84.60
Ionosphere	736	35	92.26	91.03	91.03
Pima	3068	9	73.01	72.62	72.62
TicTacToe	2396	10	91.95	90.19	92.41
WhiteHouse	439	17	98.30	97.12	97.41
xd6	604	11	75.82	75.49	75.49
Average	1728	14	85.60	84.34	85.22

Appendix: From 2-BOOST to k-BOOST

All the results we have established above aim at boosting two weak learners in parallel. Recall that the advantage of our approach is that learners collaborate and contribute to the definition of the reweighting rule, at each step. We have shown in the experiments that such an approach was more relevant than any combinaison, computed *a posteriori*, of strong hypotheses resulting of two independent (thus blind) boosting procedures.

In this section, we investigate the problem of boosting k weak learners in parallel rather that "only" two. Basically, this leads us to study Algorithm 3 below. As before, we consider a sample $\text{LS} = \{(x_1, y_1), \ldots, (x_m, y_m)\}$ drawn from a fixed but unknown distribution $\mathcal{D}$ over $\mathcal{X} \times \{-1, +1\}$. We assume that each example is described with strongly heterogeneous features, so $\mathcal{X}$ is some Cartesian product $\mathcal{X}_1 \times \mathcal{X}_2 \times \ldots \times \mathcal{X}_k$ and we assume that we have k algorithms, denoted $\text{WL}_1, \ldots, \text{WL}_k$, which will be used to learn from on their specific subset of features.

As ADABOOST and 2-BOOST, k-BOOST aims at minimizing the empirical error of the final (strong) hypothesis:

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

It is not difficult to show that minimizing this error consists in minimizing the Z_t function. Indeed,

Algorithm 3 Pseudo-code of k-BOOST.

Require: A set of weak learners $\text{WL}_1, \ldots, \text{WL}_k$,
a sample $\text{LS} = \{(x_1, y_1), \ldots, (x_m, y_m)\}$,
the maximum number T of iterations
Ensure: The (strong) combined hypothesis H_T

for $i = 1$ **to** m **do**
  $w_1(x_i) \longleftarrow 1/m$
end for
for $t = 1$ **to** T **do**
  for $j = 1$ **to** k **do**
    $h_{jt} \longleftarrow \text{WL}_j(\text{LS}, \mathbf{w}_t)$
  end for
  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 $c_{1t}, \ldots, c_{kt} \in \mathbb{R}$ **that minimizes** $Z_t(c_{1t}, \ldots, c_{kt})$
  let $Z_t = Z_t(c_{1t}, \ldots, c_{kt})$
  for $i = 1$ **to** m **do**
    $w_{t+1}(x_i) \longleftarrow w_t(x_i) \exp\left(-\sum_{j=1}^{k} c_{jt} y_i h_{jt}(x_i)\right) / Z_t$
  end for
end for
return H_T **with** $H_T(x) = \text{sign}\left(\sum_{t=1}^{T} \sum_{j=1}^{k} c_{jt} h_{jt}(x)\right)$

extending Lemma 3.1, we get:

$$\varepsilon(H_T, \text{LS}) \leq \left(\prod_{t=1}^{T} Z_t\right), \text{ where } Z_t = \sum_{i=1}^{m} w_t(x_i) \exp\left(\sum_{j=1}^{k} -c_{jt} y_i h_{jt}(x_i)\right).$$

Moreover, a global minimum of Z_t exists, because Lemma 3.2 generalizes, that is, Z_t is still a convex function. However, contrary to what happens in the case $k = 2$, an analytic expression of the optimal coefficients $c_{1t}, \ldots, c_{kt}$ that minimize Z_t cannot be found. They can only be approximated by using a standard Newton-Raphson method, for instance.

The probabilistic interpretation of Z_t as a Laplace transform (see Section 3.2) also generalizes:

$$Z_t(c_{1t}, \ldots, c_{kt}) \quad = \quad \mathbb{E}\left[\exp\left(\sum_{j=1}^{k} -c_{jt} X_j\right)\right],$$

and

$$\frac{\partial^{p_1+\ldots+p_k} Z_t}{\partial c_{1t}^{p_1} \ldots \partial c_{kt}^{p_k}}(0, \ldots, 0) \quad = \quad (-1)^{p_1+\ldots+p_k} \mathbb{E}[X_1^{p_1} \ldots X_k^{p_k}],$$

Once computed, the derivatives of Z_t and the previous relations show that $2^k - 1$ moments $\mathbb{E}[X_1^{p_1} \ldots X_k^{p_k}]$ are necessary to describe Z_t. So proving that $Z_t < 1$ under the standard weak learning assumption is clearly intricate, although probably correct.

At last, concerning the generalization error, the analysis of the penalty term still holds, but of course, showing that the margin-error is < 1 is impossible using Schapire & Freund's standard technique [10].

References

[1] Breiman, L.: Bagging Predictors, *Machine Learning*, **24**(2), 1996, 123–140.

[2] Breiman, L.: Random Forests, *Machine Learning*, **45**(1), 2001, 5–32.

[3] Burges, C. J. C.: A Tutorial on Support Vector Machines for Pattern Recognition, *Data Mining and Knowledge Discovery*, **2**, 1998, 121–167.

[4] Callut, J., Dupont, P.: Inducing Hidden Markov Models to Model Long-Term Dependencies, *Proc. of the 16th European Conference on Machine Learning (ECML'05)*, LNAI 3720, 2005.

[5] Cherkauer, K. J.: Human Expert-Level Performance on a Scientific Image Analysis Task by a System Using Combined Artificial Neural Networks, *Working Notes, Integrating Multiple Learned Models for Improving and Scaling Machine Learning Algorithms Workshop, 13th National Conference on Artificial Intelligence*, 1996.

[6] Cristianini, N., Shawe-Taylor, J.: *An Introduction to Support Vector Machines and other Kernel-based Learning Methods*, Cambridge University Press, 2000.

[7] Denis, F., Esposito, Y., Habrard, H.: Learning Rational Stochastic Languages, *Proc. of the 19th Conference on Computational Learning Theory (COLT'06)*, LNAI 4005, 2006.

[8] Dietterich, T. G.: Ensemble Methods in Machine Learning, *Proc. of the 1st International Workshop on Multiple Classifier Systems*, LNCS 1857, 2000.

[9] Durbin, R., Eddy, S. R., Krogh, A., Mitchison, G.: *Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids*, Cambridge University Press, 1999.

[10] Freund, Y., Schapire, R. E.: Experiments with a New Boosting Algorithms, *Proc. of the 13th International Conference on Machine Learning (ICML'96)*, 1996.

[11] Freund, Y., Schapire, R. E.: A Decision-Theoretic Generalization of Online Learning and an Application to Boosting, *Journal of Computer and System Sciences*, **55**(1), 1997, 119–139.

[12] Gama, J., Brazdil, P.: Cascade Generalization, *Machine Learning*, **41**(3), 2000, 315–343.

[13] Garcia-Salicetti, S., Beumier, C., Chollet, G., Dorizzi, B., Leroux-Les-Jardins, J., Lunter, J., Ni, Y., Petrovska-Delacretaz, D.: BIOMET: A Multimodal Person Authentication Database Including Face, Voice, Fingerprint, Hand and Signature Modalities, *Proc. of the 4th International Conference on Audio and Video-Based Biometric Person Authentication (AVBPA'03)*, LNCS 2688, 2003.

[14] Goodman, J.: *A Bit of Progress in Language Modeling*, Technical Report MSR-TR-2001-72, Microsoft Research, 2001.

[15] de la Higuera, C.: A Bibliographic Survey on Grammatical Inference, *Pattern Recognition*, **38**(9), 2005, 1332–1348.

[16] Kearns, M. J., Vazirani, U. V.: *An Introduction to Computational Learning Theory*, M.I.T. Press, 1994.

[17] Kohavi, R.: A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection, *Proc. of the 15th International Joint Conference on Artificial Intelligence (IJCAI'95)*, 1995.

[18] Koltchinskii, V., Panchenko, D.: Empirical Margin Distributions and Bounding the Generalization Error of Combined Classifiers, *Annals of Statistics*, **30**(1), 2002, 1–50.

[19] Meir, R., Raetsch, G.: An Introduction to Boosting and Leveraging, *Advanced Lectures on Machine Learning*, LNAI 2600, 2003.

[20] Schapire, R. E., Freund, Y., Bartlett, P., Lee, W. S.: Boosting the Margin: A New Explanation for the Effectiveness of Voting Methods, *Annals of Statistics*, **26**, 1998, 1651–1686.

[21] Schapire, R. E., Singer, Y.: Improved Boosting Algorithms using Confidence-rated Predictions, *Proc. of the 11th International Conference on Computational Learning Theory (COLT'98)*, 1998.

[22] Vapnik, V.: *Statistical Learning Theory*, John Wiley, 1998.

[23] Wolpert, D. H.: Stacked Generalization, *Neural Networks*, **5**(2), 1992, 241–259.