– Supplementary Material –
From Cost-Sensitive Classification to Tight F-measure Bounds

Kevin Bascol [1,3], Rémi Emonet[1], Elisa Fromont[2], Amaury Habrard[1], Guillaume Metzler[1,4], and Marc Sebban[1]

[1] Laboratoire Hubert Curien UMR 5516, Univ Lyon, UJM-Saint-Etienne, F-42023, Saint-Etienne,
[2]IRISA/Inria, Univ. Rennes 1, 35042 Rennes cedex, France
[3] Bluecime inc., France.
[4] Blitz Business Service inc., France.

The goal of this document is to:

- detail the proof of the results provided in the main article,
- develop the multi-class extension,
- provide illustrations and results on all considered datasets,
- give numerical values used to plot the curves (for easier reproducibility).

For the sake of clarity, we will remind each statement before giving its proof. We also recall the notations and the definitions that are used for our purpose.
In the body of the paper, the error profile of an hypothesis h as been defined as $\mathbf{E}(h) = (e_1(h), e_2(h)) = (FN(h), FP(h))$. In the binary setting and using the previous notations, the F-Measure is defined by:

$$F(\boldsymbol{e}) = \frac{(1+\beta^2)(P-e_1)}{(1+\beta^2)P - e_1 + e_2}. \tag{1}$$

1 Main results of the article

In this section, we provide all the proofs of the main article but only in the binary setting.

1.1 Pseudo-linearity of F-Measure

We aim to prove the following proposition, which plays a key role to provide a the bound on the F-measure.

Proposition 1. *The F-measure, F, is a pseudo-linear function.*

Proof. We need to show that both F and $-F$ are pseudo-convex, i.e. that we have:

$$\langle \nabla F(\boldsymbol{e}), (\boldsymbol{e}' - \boldsymbol{e}) \rangle \geq 0 \implies F(\boldsymbol{e}') \geq F(\boldsymbol{e}). \tag{2}$$

The gradient of the F-measure is defined by:

$$\nabla F(\boldsymbol{e}) = -\frac{1+\beta^2}{((1+\beta^2)P - e_1 + e_2)^2} \begin{pmatrix} \beta^2 P + e_2 \\ P - e_1 \end{pmatrix}.$$

We now develop the left hand side of the implication (2):

$$\begin{aligned} \langle \nabla F(\boldsymbol{e}), (\boldsymbol{e}' - \boldsymbol{e}) \rangle &\geq 0, \\ -\frac{1+\beta^2}{((1+\beta^2)P - e_1 + e_2)^2} \left[(\beta^2 P + e_2)(e_1' - e_1) + (P - e_1)(e_2' - e_2) \right] &\geq 0, \end{aligned}$$

so,

$$\begin{aligned} -(\beta^2 P + e_2)(e_1' - e_1) - (P - e_1)(e_2' - e_2) &\geq 0, \\ -\beta^2 P(e_1' - e_1) - e_1' e_2 + e_1 e_2 + P(e_2 - e_2') + e_1 e_2' - e_1 e_2 &\geq 0, \\ -\beta^2 P(e_1' - e_1) + P(e_2 - e_2') + e_1 e_2' - e_1' e_2 &\geq 0, \\ -\beta^2 P e_1' + \beta^2 P e_1 + P e_2 - P e_2' + e_1 e_2' - e_1' e_2 &\geq 0. \end{aligned}$$

so

$$-\beta^2 P e_1' + P e_2 - e_1' e_2 \;\geq\; -\beta^2 P e_1 + P e_2' - e_1 e_2'.$$

Now we add $-P(e_1 + e_1')$ on both side of the inequality, so we have:

$$\begin{aligned} -\beta^2 P e_1' + P e_2 - e_1' e_2 - P(e_1 + e_1') &\geq -\beta^2 P e_1 + P e_2' - e_1 e_2' - -P(e_1 + e_1'), \\ -(1+\beta^2) P e_1' + P e_2 - e_1' e_2 - P e_1 &\geq -(1+\beta^2) P e_1 + P e_2' - e_1 e_2' - P e_1'. \end{aligned}$$

Then, we add $e_1 e_1'$ on both sides:

$$\begin{aligned} -(1+\beta^2) P e_1' + P e_2 - e_1' e_2 - P e_1 + e_1 e_1' &\geq -(1+\beta^2) P e_1 + P e_2' - e_1 e_2' - P e_1' + e_1 e_1', \\ -(1+\beta^2) P e_1' - (P - e_1') e_1 + (P - e_1') e_2 &\geq -(1+\beta^2) P e_1 - (P - e_1) e_1' + (P - e_1) e_2'. \end{aligned}$$

Finally, by adding $(1+\beta^2)P^2$ on both sides of the inequality and factorizing with the terms $-(1+\beta^2)Pe_1'$ on the left (respectively $-(1+\beta^2)Pe_1$ on the right), we get:

$$\begin{aligned} (1+\beta^2)P(P - e_1') - (P - e_1')e_1 + (P - e_1')e_2 &\geq (1+\beta^2)P(P - e_1) - (P - e_1)e_1' + (P - e_1)e_2', \\ (1+\beta^2)P(P - e_1') - (P - e_1')e_1 + (P - e_1')e_2 &\geq (1+\beta^2)P(P - e_1) - (P - e_1)e_1' + (P - e_1)e_2', \\ (P - e_1')((1+\beta^2)P - e_1 + e_2) &\geq (P - e_1)((1+\beta^2)P e_1' + e_2'), \\ (1+\beta^2)(P - e_1')((1+\beta^2)P - e_1 + e_2) &\geq (1+\beta^2)(P - e_1)((1+\beta^2)P e_1' + e_2'), \\ \frac{(P - e_1')}{(1+\beta^2)P - e_1' + e_2'} &\geq \frac{(P - e_1)}{(1+\beta^2)P - e_1 + e_2}, \\ \frac{(1+\beta^2)(P - e_1')}{(1+\beta^2)P - e_1' + e_2'} &\geq \frac{(1+\beta^2)(P - e_1)}{(1+\beta^2)P - e_1 + e_2}, \\ F(\boldsymbol{e}') &\geq F(\boldsymbol{e}). \end{aligned}$$

The proof is similar for $-F$.
We have shown that both F and $-F$ are pseudo-convex so F is pseudo-linear. □

We can now use this property to derive our bound. However, we have seen that the bound still depends on to other parameters M_{min} and M_{max} that we should compute.

1.2 Computation of the values of M_{min} and M_{max}.

We aim to show how we can solve the optimization problems that define M_{min} and M_{max} and show how it can be reduced to a simple convex optimization problem where the set of constraints is a convex polygon.

Computation of M_{max}

Now, we would like to give an explicit value for M_{max}. This value can be obtained by solving the following optimization problem:

$$\max_{\boldsymbol{e'} \in \mathcal{E}(\mathcal{H})} e'_2 - e'_1 \quad s.t. \quad F_\beta(\boldsymbol{e'}) > F_\beta(\boldsymbol{e}).$$

In the binary case, setting $\boldsymbol{e} = (e_1, e_2)$ and $\boldsymbol{e'} = (e'_1, e'_2)$. We can write $F_\beta(\boldsymbol{e'}) > F_\beta(\boldsymbol{e})$ as:

$$\frac{(1+\beta^2)(P - e'_1)}{(1+\beta^2)P - e'_1 + e'_2} > \frac{(1+\beta^2)(P - e_1)}{(1+\beta^2)P - e_1 + e_2},$$

Now we develop and reduce these expressions.

$$\begin{aligned}
(P - e'_1)[(1+\beta^2)P - e_1 + e_2] &> (P - e_1)[(1+\beta^2)P - e'_1 + e'_2]), \\
(1+\beta^2)P^2 - (1+\beta^2)Pe'_1 + (P - e'_1)(e_2 - e_1) &> (1+\beta^2)P^2 - (1+\beta^2)Pe_1 + (P - e_1)(e'_2 - e'_1), \\
(1+\beta^2)P(e_1 - e'_1) + P(e_2 - e_1 + e'_1 - e'_2) &> e_2e'_1 - e_1e'_2 + e'_1e_1 - e_1e'_1.
\end{aligned}$$

Now, we set: $e'_1 = e_1 + \alpha_1$ and $e'_2 = e_2 + \alpha_2$. In other words, we study how much we have to change $\boldsymbol{e}'$ from $\boldsymbol{e}$ to solve our problem. We can then write:

$$\begin{aligned}
-(1+\beta^2)P\alpha_1 + P(\alpha_1 - \alpha_2) &> e_2(e_1 + \alpha_1) - e_1(e_2 + \alpha_2), \\
\alpha_1(-(1+\beta^2)P + P - e_2) + \alpha_2(-P + e_1) &> 0, \\
\alpha_1(\beta^2 P + e_2) &< -\alpha_2(P - e_1).
\end{aligned}$$

Thus, the optimization problem can be rewritten as:

$$\begin{aligned}
\max_{\boldsymbol{\alpha}} \quad & \alpha_2 - \alpha_1, \\
s.t. \quad & \alpha_1 < \frac{-\alpha_2(P - e_1)}{\beta^2 P + e_2}, \\
& \alpha_1 \in [-e_1, P - e_1], \\
& \alpha_2 \in [-e_2, N - e_2].
\end{aligned}$$

The optimization problem consists of maximizing a difference under a polygon set of constraints. In the binary setting, the set of constraints can be represented as shown in Fig. 1 where the line $\mathcal{D}$ is defined by the following equation:

$$\alpha_1 = \frac{-\alpha_2(P - e_1)}{\beta^2 P + e_2}. \tag{3}$$

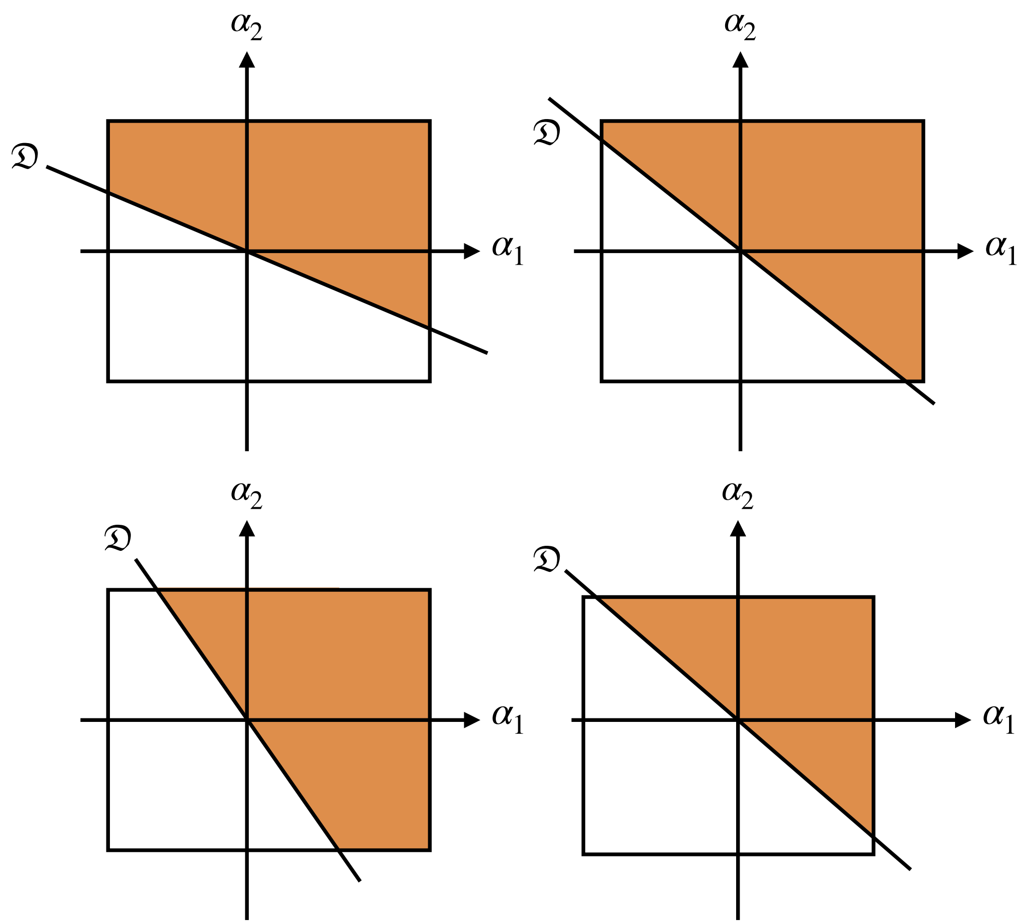


Figure 1: Geometric representation of the optimization problem. The rectangle represents the constraint $(\alpha_2, \alpha_1) \in [-e_2, N - e_2] \times [e_1, P - e_1]$. The white area represents the set of value (α_2, α_1) for which the inequality constraint holds. the four figures represent the four possibility for the position of the line $\mathcal{D}$ on the rectangle. See the computation of M_{min} to see that cases represented by the two figures at the bottom never happen.

To maximize the difference, we should maximize the value of α_2 and minimize the value of α_1, i.e. the solution is located in the bottom right region of each figure. A quick study of these figures shows that the lowest value of α_1 we can reach is $-e_1$.

We shall now study where the line $\mathcal{D}$ intersects the rectangle to have the solution with respect to α_2. If $\mathcal{D}$ does not intersect the line of equation $\alpha_1 = -e_1$ in the rectangle (i.e. it intersects with the right side of the rectangle) then $\alpha_2 = N - e_2$. Else, it intersects with the bottom face of the rectangle, then we determine the value of α_2 using Eq. (3) and $\alpha_2 = \dfrac{(\beta^2 P + e_2)e_1}{P - e_1}$.

Finally, the solution of the optimization problem is:

$$(\alpha_1, \alpha_2) = \left(-e_1, min\left(N - e_2, \frac{(\beta^2 P + e_2)e_1}{P - e_1}\right)\right),$$

and the optimal value M_{max} is defined by:

$$M_{max} = e_2 + \min\left(N - e_2, \frac{(\beta^2 P + e_2)e_1}{P - e_1}\right).$$

Computation of M_{min}

We now aim to solve the following optimization problem:

$$\min_{\boldsymbol{e'} \in \mathcal{E}(\mathcal{H})} \; e'_2 - e'_1 \quad s.t. \quad F_\beta(\boldsymbol{e'}) > F_\beta(\boldsymbol{e}).$$

As it has been done and using the same notations as in the previous section, we can rewrite the optimization problem as follows:

$$\begin{aligned}
\min_{\boldsymbol{\alpha}} \quad & \alpha_2 - \alpha_1, \\
s.t. \quad & \alpha_1 < \frac{-\alpha_2(P - e_1)}{\beta^2 P + e_2}, \\
& \alpha_1 \in [-e_1, P - e_1], \\
& \alpha_2 \in [-e_2, N - e_2].
\end{aligned}$$

The constraints remain unchanged. However, to minimize this difference, we have to maximize the value of α_1 and minimize the value of α_2, i.e. we are interested in the upper left region of each rectangles. In each cases represented in Fig 1, we see that the minimum of α_2 is equal to $-e_2$.

If we have a look at the two figures at the bottom of Fig. 1, we see that the optimal value of α_1 is equal to $P - e_1$. However, this value is not in the image of the function of α_2 defined by Eq (3). In fact, according to Eq. (3), the image of $\alpha_2 = -e_2$ is equal to $\frac{e_2(P - e_1)}{\beta^2 P + e_2}$ which is lower than $P - e_1$. So the two figures at the bottom represent cases that never happen and the intersection of $\mathcal{D}$ with the rectangle of constraint is on left part of the rectangle.

Finally, the solution of the optimization problem is:

$$(\alpha_1, \alpha_2) = \left(\frac{e_2(P - e_1)}{\beta^2 P + e_2}, -e_2\right),$$

and the optimal value M_{min} is defined by:

$$M_{min} = -e_1 - \frac{e_2(P - e_1)}{\beta^2 P + e_2}.$$

Now that we have provided all the details to compute and plot our bound, it remains to explain how to compute the bound from Parambath et al. (2014) with respect to any cost parameters t, t' for a fair comparison.

1.3 Rewriting the bound of Parambath et al. (2014)

For the sake of clarity we restate the Proposition 5 of Parambath et al. (2014) for our purpose:

Proposition 2. *Let $t, t \in [0,1]$ and $\varepsilon_1 \geq 0$. Suppose that it exists $\Phi > 0$ such that for all $\boldsymbol{e}, \boldsymbol{e}' \in \mathcal{E}(\mathcal{H})$ satisfying $F(\boldsymbol{e}') > F(\boldsymbol{e})$, we have:*

$$F(\boldsymbol{e}') - F(e) \geq \Phi\langle \boldsymbol{a}(t'), \boldsymbol{e} - \boldsymbol{e}'\rangle. \tag{4}$$

Furthermore, suppose that we have the two following conditions

$$(i)\ \ \|\boldsymbol{a}(t) - \boldsymbol{a}(t')\|_2 \leq 2|t - t'| \qquad (ii)\ \ \langle \boldsymbol{a}(t), \boldsymbol{e}\rangle \leq \underset{\mathbf{e}'' \in \mathcal{E}(\mathcal{H})}{min} \langle \boldsymbol{a}(t), \mathbf{e}''\rangle + \varepsilon_1$$

Let us also set $M = \underset{\mathbf{e}'' \in \mathcal{E}(\mathcal{H})}{max} \|\boldsymbol{e}''\|_2$, then we have:

$$F(\boldsymbol{e}') \leq F(\boldsymbol{e}) + \Phi\varepsilon_1 + 4M\Phi|t' - t|.$$

According to the authors, the point (i) is a consequence of a of being Lipschitz continous with Lipschtiz constant equal to 2. The point (ii) is just the expression of the sub-optimality of the learned classifier.

Proof. For all $\boldsymbol{e}, \tilde{\boldsymbol{e}} \in \mathcal{E}(\mathcal{H})$ and $t, t' \in [0,1]$, we have:

$$\begin{aligned} \langle \boldsymbol{a}(t), \tilde{\boldsymbol{e}}\rangle &= \langle \boldsymbol{a}(t) - \boldsymbol{a}(t'), \tilde{\boldsymbol{e}}\rangle + \langle a(t'), \tilde{\boldsymbol{e}}\rangle, \\ &\leq \langle \boldsymbol{a}(t'), \tilde{\boldsymbol{e}}\rangle + 2M|t' - t|. \end{aligned}$$

Where we have successively applied the Cauchy-Schwarz inequality and (i). Then:

$$\underset{\mathbf{e}'' \in \mathcal{E}(\mathcal{H})}{min} \langle \boldsymbol{a}(t), \mathbf{e}''\rangle \leq \underset{\mathbf{e}'' \in \mathcal{E}(\mathcal{H})}{min} \langle \boldsymbol{a}(t'), \mathbf{e}''\rangle + 2M|t' - t| = \langle \boldsymbol{a}(t'), \boldsymbol{e}'\rangle + 2M|t' - t|, \tag{5}$$

where $\boldsymbol{e}'$ denote the error profile learned by the optimal classifier trained with the cost function $\boldsymbol{a}(t')$ and is such that $F(\boldsymbol{e}') > F(\boldsymbol{e})$. Then, writing $\langle \boldsymbol{a}(t'), \boldsymbol{e}\rangle = \langle \boldsymbol{a}(t') - \boldsymbol{a}(t), \boldsymbol{e}\rangle + \langle \boldsymbol{a}(t), \boldsymbol{e}\rangle$ and applying the Cauchy-Schwarz inequality, we have:

$$\begin{aligned} \langle \boldsymbol{a}(t'), \boldsymbol{e}\rangle &\leq \langle \boldsymbol{a}(t), \boldsymbol{e}\rangle + 2M|t' - t|, \\ &\leq \underset{\mathbf{e}'' \in \mathcal{E}(\mathcal{H})}{min} \langle \boldsymbol{a}(t), \mathbf{e}''\rangle + \varepsilon_1 + 2M|t' - t|, \\ &\leq \langle \boldsymbol{a}(t'), \boldsymbol{e}'\rangle + \varepsilon_1 + 4M|t' - t|, \end{aligned}$$

where the second inequality comes from (ii) and the last inequality comes from Eq. (5). By plugging this last inequality in inequality (4), we get the result.
Furthermore, the existence of the constant Φ has been proved by the authors and is equal to $(\beta^2 P)^{-1}$ □

Remark. This bound can be used in both binary and multi-class setting.

2 The multi-class setting

For a given hypothesis $h \in \mathcal{H}$ learned from $\mathbf{X}$, the errors that h makes can be summarized in an error profile defined as $\mathbf{E}(h) \in \mathbb{R}^{2L}$:

$$\mathbf{E}(h) = (\mathtt{FN}_1(h), \mathtt{FP}_1(h), ..., \mathtt{FN}_L(h), \mathtt{FP}_L(h)),$$

where $\mathtt{FN}_i(h)$ (resp. $\mathtt{FP}_i(h)$) is the proportion of False Negative (resp. False Positive) that h yields for class i.
In a multi-class setting with L classes P_k, $k = 1, ..., L$ denotes the proportion of examples in class k and $\boldsymbol{e} = (e_1, e_2, ..., e_{2L-1}, e_{2L})$ denotes the proportions of misclassified examples composing the error profile.
The multi-class-micro F-measure, $mcF(\boldsymbol{e})$ with L classes is defined by:

$$mcF(\boldsymbol{e}) = \frac{(1+\beta^2)(1 - P_1 - \sum_{k=2}^{L} e_{2k-1})}{(1+\beta^2)(1-P_1) - \sum_{k=2}^{L} e_{2k-1} + e_1}.$$

In this section, we aim to derive all the results presented in the binary case in a multi-class setting.

2.1 Pseudo-linearity

Proposition 3. *The multi-class-micro F-measure, mcF, is a pseudo-linear function with respect to $\boldsymbol{e}$.*

Proof. As in the binary cases, we have to prove that both mc The gradient of the multi-class-micro F-measure,mcF_β, is defined by:

$$\nabla mcF(\boldsymbol{e}) = \frac{-(1+\beta^2)}{(1+\beta^2)(1-P_1) - \sum_{k=2}^{L} e_{2k-1} + e_1} \begin{cases} 1 - P_1 - \sum_{k=2}^{L} e_{2k-1} & \text{w.r.t. } e_1, \\ \beta^2(1-P_1) + e_1 & \text{w.r.t. } e_k \ \forall k = 2, ..., L. \end{cases}$$

The proof is similar to the proof of Proposition 1. The scheme is the same, we simply have to do the following changes of notation in the proof:

$$\begin{aligned} e_1 &\leftarrow \sum_{k=2}^{L} e_{2k-1}, \\ e_2 &\leftarrow e_1, \\ P &\leftarrow 1 - P_1. \end{aligned}$$

□

2.2 Derivation of the bound

As it was done in the binary case, we will use the property of pseudo-linearity of $mcF(\boldsymbol{e})$ to bound the difference of micro F-measure in terms of the parameters of our weighted function. First, we introduce the definition of our weighted function $\boldsymbol{a} : \mathbb{R} \to \mathbb{R}^{2L}$ and express the difference of micro F-measure of two error profiles in function of the two error profiles.
In this section, for the sake of clarity, we will set $\hat{e} = \sum_{k=2}^{L} e_{2k-1}$.

First step: impact of a change in the error profile

Using the property of pseudo-linearity, we can show that it exists two functions $\boldsymbol{a} : \mathbb{R} \to \mathbb{R}^{2L}$ and $b : \mathbb{R} \to \mathbb{R}$ defined by:

$$0 = \langle \boldsymbol{a}(mcF(\boldsymbol{e})), \boldsymbol{e}\rangle + b(mcF(\boldsymbol{e})),$$

where:

$$\boldsymbol{a}(t) = \begin{cases} 1+\beta^2 - t & \text{for } e_{2k-1},\ k = 2, ..., L \\ t & \text{for } e_1, \\ 0 & \text{otherwise,} \end{cases} \quad \text{and} \quad b(t) = (t-1)(1+\beta^2)(1-P_1).$$

From these definitions we can write:

$$\begin{aligned} \langle \boldsymbol{a}(mcF(\boldsymbol{e}')), \boldsymbol{e} - \boldsymbol{e}'\rangle &= \langle \boldsymbol{a}(mcF(\boldsymbol{e}')), \boldsymbol{e}\rangle + b(mcF(\boldsymbol{e}')), \\ &= \langle \boldsymbol{a}(mcF(\boldsymbol{e}')) - \boldsymbol{a}(mcF(\boldsymbol{e})), \boldsymbol{e}\rangle - b(mcF(\boldsymbol{e})) + b(mcF(\boldsymbol{e}')), \\ &= (mcF(\boldsymbol{e}') - mcF(\boldsymbol{e}))(1+\beta^2)(1-P_1) \\ &\quad + (mcF(\boldsymbol{e}') - mcF(\boldsymbol{e}))e_1 + (mcF(\boldsymbol{e}) - mcF(\boldsymbol{e}'))\hat{e}, \\ &= (mcF(\boldsymbol{e}') - mcF(\boldsymbol{e}))\left((1+\beta^2)(1-P_1) - \hat{e} + e_1\right). \end{aligned}$$

We can now write the difference of micro-F-measure as:

$$mcF(\boldsymbol{e}') - mcF(\boldsymbol{e}) = \Phi_{\boldsymbol{e}} \cdot \langle \boldsymbol{a}(t), \boldsymbol{e} - \boldsymbol{e}'\rangle,$$

where:

$$\Phi_{\boldsymbol{e}} = \frac{1}{(1+\beta^2)(1-P_1) - \hat{e} + e_1},$$

Second step: a bound on the micro F-measure $mcF(\boldsymbol{e})$

We suppose that we have a value of t for which a weighted-classifier with weights $a(t)$ has been learned. This classifier has an error profile $\boldsymbol{e}$ and a F-measure $mcF(\boldsymbol{e})$. We now imagine a hypothetical classifier that is learned with weights $a(t')$, and we denote by $\boldsymbol{e}'$ the error profile of this classifier. For any value of t', we derive an upper bound on the on the F-measure $mcF(\boldsymbol{e}')$ that this hypothetical classifier can achieve.

$$\begin{aligned} mcF(\boldsymbol{e}') - mcF(\boldsymbol{e}) &= \Phi_{\boldsymbol{e}} \cdot \langle \boldsymbol{a}(t'), \boldsymbol{e} - \boldsymbol{e}'\rangle, \\ &= \Phi_{\boldsymbol{e}} \cdot \left(\langle \boldsymbol{a}(t'), \boldsymbol{e}\rangle - \langle \boldsymbol{a}(t'), \boldsymbol{e}\rangle\right), \\ &= \Phi_{\boldsymbol{e}} \cdot \left(\langle \boldsymbol{a}(t') - \boldsymbol{a}(t), \boldsymbol{e}\rangle + \langle \boldsymbol{a}(t), \boldsymbol{e}\rangle - \langle \boldsymbol{a}(t'), \boldsymbol{e}'\rangle\right), \\ &= \Phi_{\boldsymbol{e}} \cdot \left(\langle (t'-t, t-t'), \boldsymbol{e}\rangle + \langle \boldsymbol{a}(t), \boldsymbol{e}\rangle - \langle \boldsymbol{a}(t'), \boldsymbol{e}'\rangle\right), \\ &= \Phi_{\boldsymbol{e}} \cdot \left((t'-t)(e_1 - \hat{e}) + \langle \boldsymbol{a}(t), \boldsymbol{e}\rangle - \langle \boldsymbol{a}(t'), \boldsymbol{e}'\rangle\right), \\ &\leq \Phi_{\boldsymbol{e}} \cdot \left(\langle \boldsymbol{a}(t), \boldsymbol{e}'\rangle + \varepsilon_1 - \langle \boldsymbol{a}(t'), \boldsymbol{e}'\rangle + (t'-t)(e_1 - \hat{e})\right), \\ &\leq \Phi_{\boldsymbol{e}} \cdot \left((t'-t)(e_1 - \hat{e}) + \varepsilon_1 - (t'-t)(e_1' - \hat{e}')\right), \\ &\leq \Phi_{\boldsymbol{e}}\varepsilon_1 + \Phi_{\boldsymbol{e}} \cdot (e_1 - \hat{e} - (e_1' - \hat{e}'))(t'-t). \end{aligned}$$

In the previous development, we have used the linearity of the inner product and the definition of $\boldsymbol{a}$. The first inequality uses the sub-optimality of the learned classifier. We then use the definition of the function $\boldsymbol{a}$.

As in the binary cases, the quantity $(e'_1 - \hat{e}')$ remains unknown. However, we are looking for a vector $\boldsymbol{e}'$ such that $mcF(\boldsymbol{e}') > mcF(\boldsymbol{e})$. So the last inequality becomes, if $t' < t$:

$$mcF(\boldsymbol{e}') - mcF(\boldsymbol{e}) \leq \Phi_{\boldsymbol{e}}\varepsilon_1 + \Phi_{\boldsymbol{e}}(e_2 - e_1 - M_{max})(t' - t),$$

and, if $t' > t$:

$$mcF(\boldsymbol{e}') - mcF(\boldsymbol{e}) \leq \Phi_{\boldsymbol{e}}\ \varepsilon_1 + \Phi_{\boldsymbol{e}}(e_2 - e_1 - M_{min})(t' - t).$$

2.3 Computation of M_{max} and M_{min} in a multiclass setting

To compute the value of both M_{max} and M_{min}, we use the same development as done in the binary setting. We have to search how to modify the vector $\boldsymbol{e}$ in order to improve the F-Measure and to maximize (or minimize) the difference: $e'_1 - \sum_{k=2}^{L} e'_{2k-1}$, where $\boldsymbol{e}' = \boldsymbol{e} + \boldsymbol{\alpha}$. As in the previous section, $\boldsymbol{\alpha}$ is the solution of the following optimization problem:

$$\begin{aligned}
\underset{\boldsymbol{\alpha}}{max} \quad & \alpha_1 - \sum_{k=2}^{L} \alpha_{2k-1}, \\
s.t. \quad & \alpha_1 < -\sum_{k=2}^{L} \alpha_{2k-1} \frac{\beta^2(1-P_1) + e_1}{1 - P_1 - \sum_{k=2}^{L} e_{2k-1}} \\
& \alpha_1 \in [-e_1, P_1 - e_1], \\
& \alpha_{2k-1} \in [-e_{2k-1}, P_{2k-1} - e_{2k-1}], \ \forall k = 2, ..., L.
\end{aligned}$$

Then we add the quantity $e_1 - \sum_{k=2}^{L} e_{2k-1}$ to this result to have the value M_{max}.
Similarly, we solve the following optimization problem:

$$\begin{aligned}
\underset{\boldsymbol{\alpha}}{min} \quad & \alpha_1 - \sum_{k=2}^{L} \alpha_{2k-1}, \\
s.t. \quad & \alpha_1 < -\sum_{k=2}^{L} \alpha_{2k-1} \frac{\beta^2(1-P_1) + e_1}{1 - P_1 - \sum_{k=2}^{L} e_{2k-1}} \\
& \alpha_1 \in [-e_1, P_1 - e_1], \\
& \alpha_{2k-1} \in [-e_{2k-1}, P_{2k-1} - e_{2k-1}], \ \forall k = 2, ..., L.
\end{aligned}$$

Then we add the quantity $e_1 - \sum_{k=2}^{L} e_{2k-1}$ to this result to have the value M_{min}.

3 Extended Experiments

This section is dedicated to the experiments. We provide all graphs and tables we were not able to give in the main paper and for all datasets.

3.1 Illustrations of unreachable regions

In this section we provide the unreachable regions (see Fig. 2) of both presented bounds, our vs. the one obtained from Parambath et al. (2014). As it was noticed in the main paper, our result gives a tighter bound on the optimal reachable F-measure. Moreover, we see that the more the data is

imbalanced, the tightest is our bound.
The fact that some points lie in the unreachable regions is explained by our setting. Indeed, we recall that we made the assumption that $\varepsilon_1 = 0$, i.e. we suppose that learned classifier is the optimal one, in terms of $0-1$ loss, but it is not the case in practice.

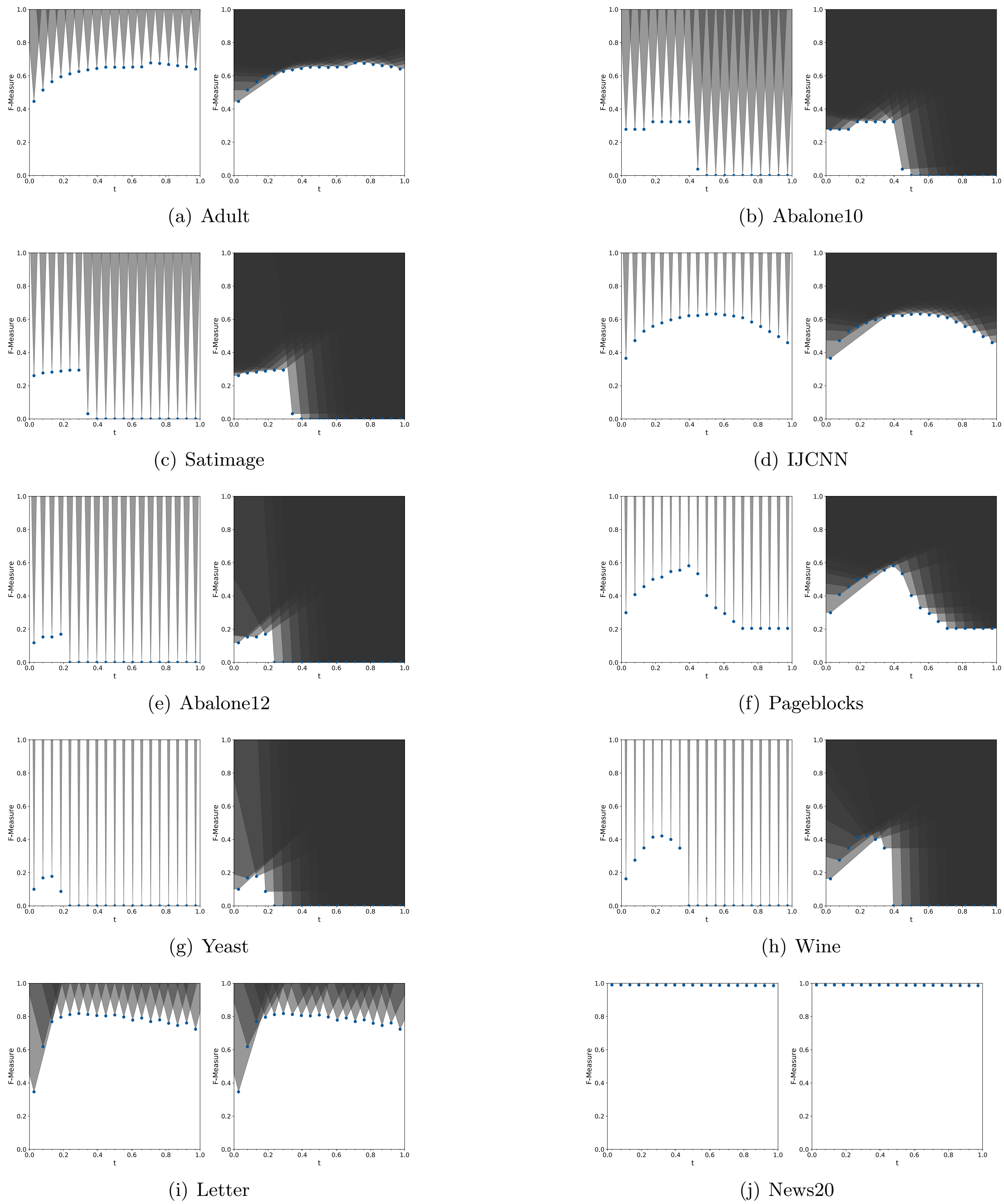


Figure 2: Unreachable regions obtained from the same 19 (t_1, F_i) points corresponding to learning weighted SVM on a grid of t values. Cones are shown for all datasets. The bound from Parambath et al. (2014) is represented on the left and our bound on the right.

3.2 Theoretical bound versus ε_1

In this section we compare our bound with the one from Parambath et al. (2014) with respect to ε_1. The graphics presented in Fig. 3 show that the bound from Parambath et al. (2014) is uninformative since the value of the best reachable F-measure is always equal to 1 except on Abalone10 dataset. We see that our bound increase mostly linearly with ε_1. the evolution is not exactly linear because the value of Φ_e depends on the error profile, so it depends on the value of the parameter t in our cost function $\boldsymbol{a}$. Note that the best classifier reaches a best F-measure in some cases (on Letter dataset for instance) which emphasize the need to look for an estimation of ε_1.

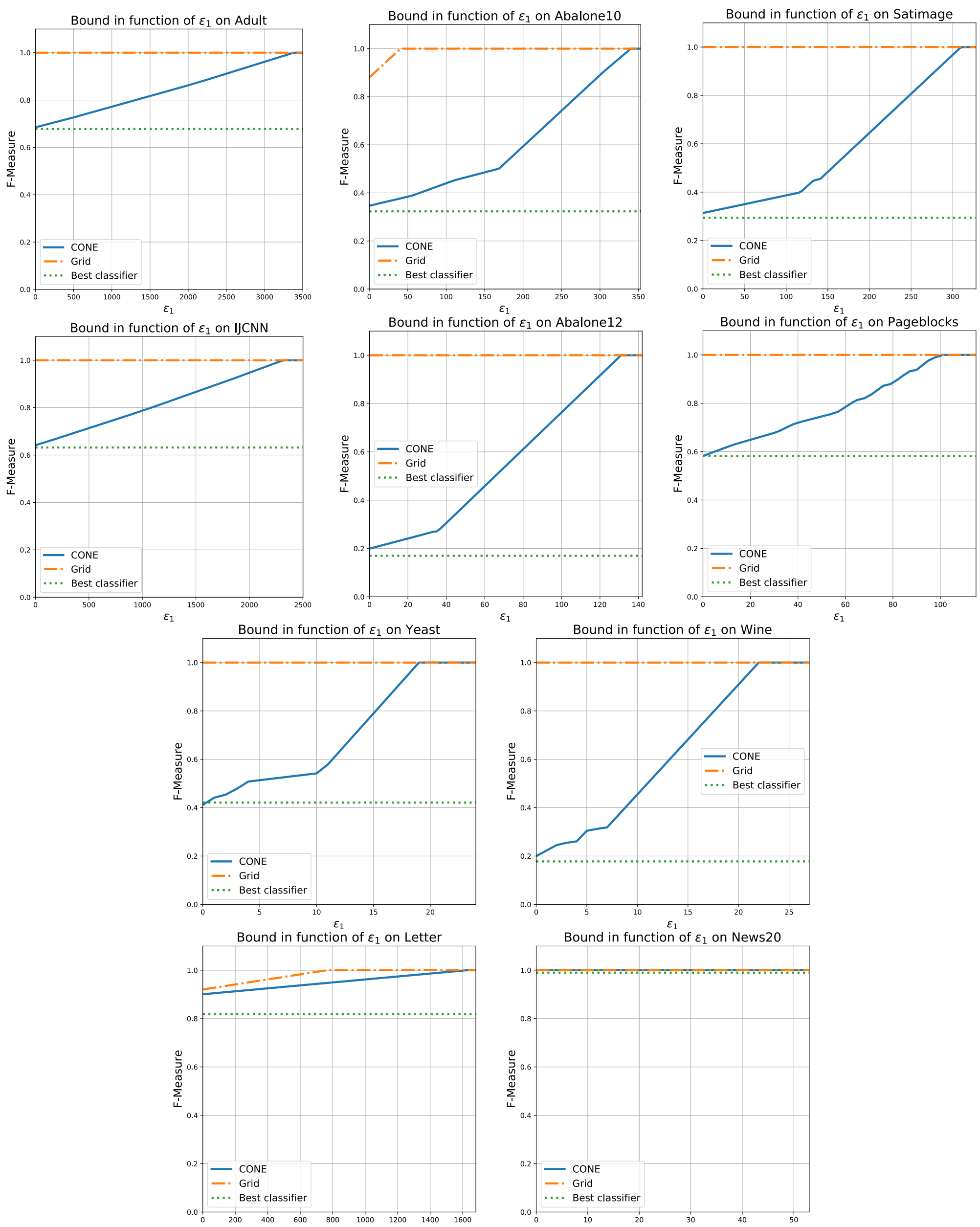
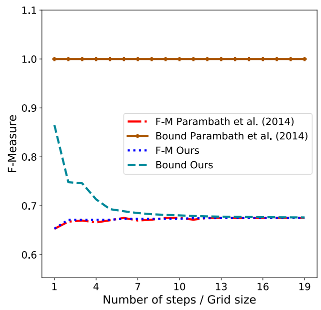
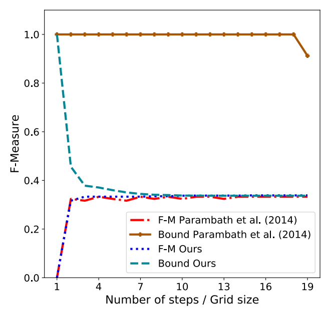


Figure 3: Bounds on the F-measure as a function of ε_1, the unknown sub-optimality of the SVM learning algorithm. Results are given on all datasets.

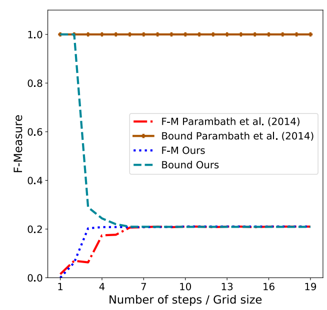
3.3 Evolution of Bounds vs. iterations/grid size



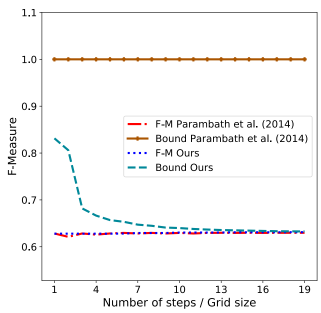
(a) Adult



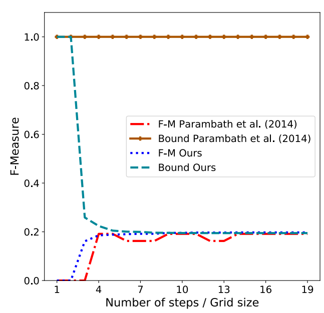
(b) Abalone10



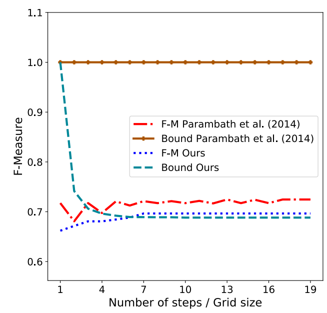
(c) Satimage



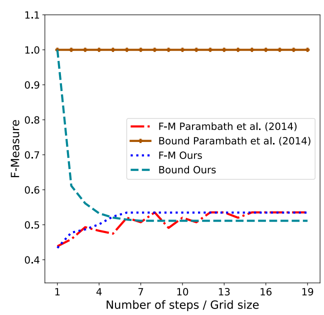
(d) IJCNN



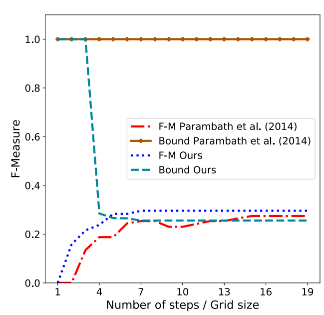
(e) Abalone12



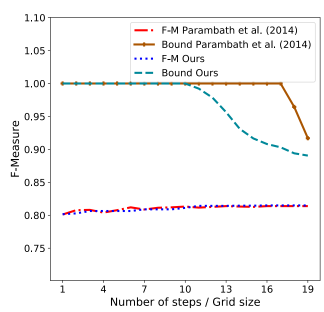
(f) Pageblocks



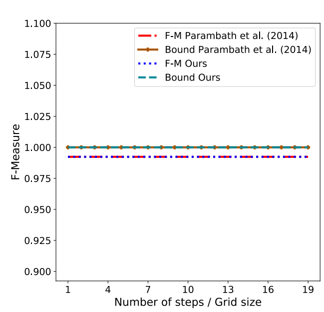
(g) Yeast



(h) Wine



(i) Letter



(j) News20

Figure 4: Comparison of our bound and the one from Parambath et al. (2014) with respect to the number of iteration/ the size of the grid. We also represent the evolution of both associated algorithms.

3.4 Test-time results and result tables of results

For the sake of clarity, only a small number of algorithms have been chosen to be represented graphically in Fig. 5.

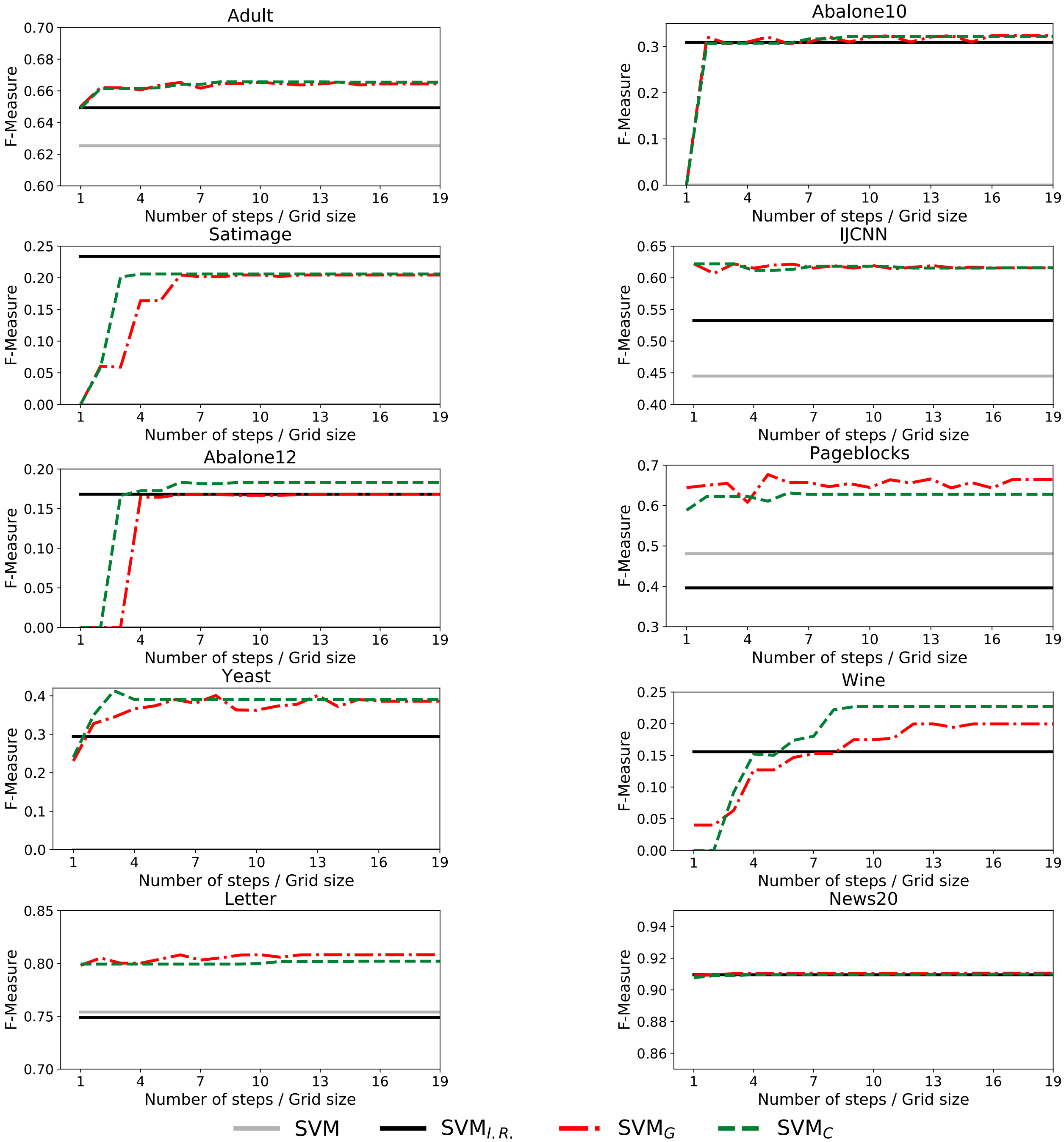


Figure 5: F-measure value with respect to the number of iterations or the size of the grid of four different algorithms, all of them are based on SVM.

To complete the results given in the main article, we provide two tables below. Table 1 gives the value of the F-measure for all experiments with SVM or Logisitic Regression based algorithms. Because we compare our method to some which uses a threshold to predict the class of an example (Narasimhan et al., 2015; Koyejo et al., 2014), we also provide a thresholded version of all algorithms in Table 2.

Table 1: Classification F-Measure for $\beta = 1$ with SVM algorithm. SVM$_G$ are reproduced from (Parambath et al., 2014) and the subscript $_{I.R.}$ is used for the classifiers trained with a cost depending on the Imbalance Ratio. The subscript $_B$ corresponds to the bisection algorithm presented in (Narasimhan et al., 2015). Finally the $_C$ stand for our wrapper CONE. The presented values are obtained by taking the mean F-Measure over 5 experiments (standard deviation between brackets).

Dataset	SVM	SVM$_{I.R.}$	SVM$_G$	SVM$_C$	LR	LR$_{I.R.}$	LR$_B$	LR$_G$	LR$_C$
Adult	62.5 (0.2)	64.9 (0.3)	66.4 (0.1)	66.5 (0.1)	63.1 (0.1)	66.0 (0.1)	66.6 (0.1)	66.5 (0.1)	66.5 (0.1)
Abalone10	0.0 (0.0)	30.9 (1.2)	32.4 (1.3)	32.2 (0.8)	0.0 (0.0)	31.9 (1.4)	31.6 (0.6)	31.7 (0.7)	30.9 (1.9)
Satimage	0.0 (0.0)	23.4 (4.3)	20.4 (5.3)	20.6 (5.6)	0.5 (0.9)	24.2 (5.3)	21.4 (4.6)	20.7 (4.8)	20.5 (5.0)
IJCNN	44.5 (0.4)	53.3 (0.4)	61.6 (0.6)	61.6 (0.6)	46.2 (0.3)	51.6 (0.3)	59.2 (0.3)	58.2 (0.2)	58.2 (0.3)
Abalone12	0.0 (0.0)	16.8 (2.7)	16.8 (4.2)	18.3 (3.3)	0.0 (0.0)	18.0 (3.5)	17.7 (3.7)	17.2 (3.1)	18.4 (2.3)
Pageblocks	48.1 (5.8)	39.6 (4.7)	66.4 (3.2)	62.8 (3.9)	48.6 (3.3)	42.4 (5.2)	55.7 (5.7)	62.8 (8.2)	59.4 (7.5)
Yeast	0.0 (0.0)	29.4 (2.9)	38.6 (7.1)	39.0 (7.5)	2.5 (5.0)	29.0 (3.5)	35.4 (15.6)	39.1 (10.1)	39.5 (9.3)
Wine	0.0 (0.0)	15.6 (5.2)	20.0 (6.4)	22.7 (6.0)	0.0 (0.0)	14.6 (3.2)	18.3 (7.2)	18.7 (4.5)	21.1 (5.2)
Letter	75.4 (0.7)	74.9 (0.8)	80.8 (0.5)	81.0 (0.4)	82.9 (0.3)	82.9 (0.3)	74.9 (0.5)	82.9 (0.2)	82.9 (0.3)
News20	90.9 (0.1)	91.0 (0.2)	91.1 (0.1)	91.0 (0.1)	90.6 (0.1)	90.6 (0.1)	89.4 (0.2)	90.6 (0.2)	90.6 (0.2)
Average	32.1 (0.7)	44.0 (2.3)	49.5 (2.9)	49.6 (2.8)	33.4 (1.0)	45.1 (2.3)	47.0 (3.9)	48.8 (3.2)	48.8 (3.2)

Table 2: Classification F-Measure for $\beta = 1$ with **thresholded** SVM algorithm. SVM$_G$ are reproduced from (Parambath et al., 2014) and the subscript $_{I.R.}$ is used for the classifiers trained with a cost depending on the Imbalance Ratio. The subscript $_B$ corresponds to the bisection algorithm presented in (Narasimhan et al., 2015). Finally the $_C$ stand for our wrapper CONE. The presented values are obtained by taking the mean F-Measure over 5 experiments (standard deviation between brackets).

Dataset	SVM	SVM$_{I.R.}$	SVM$_G$	SVM$_C$	LR	LR$_{I.R.}$	LR$_G$	LR$_C$
Adult	65.6 (0.3)	66.1 (0.2)	66.4 (0.2)	66.4 (0.1)	66.5 (0.1)	66.5 (0.1)	66.5 (0.1)	66.5 (0.1)
Abalone10	27.8 (1.2)	30.7 (2.0)	31.9 (0.6)	31.8 (1.9)	30.8 (2.2)	30.7 (1.9)	30.7 (1.9)	30.8 (2.1)
Satimage	26.7 (4.9)	29.2 (2.6)	31.6 (1.7)	30.9 (2.0)	21.2 (11.1)	28.6 (1.9)	25.3 (12.7)	25.6 (12.8)
IJCNN	63.2 (0.6)	57.4 (0.3)	62.4 (0.5)	62.6 (0.4)	59.4 (0.5)	56.5 (0.3)	59.3 (0.4)	59.3 (0.2)
Abalone12	10.2 (3.6)	16.6 (2.7)	14.5 (3.2)	16.3 (3.0)	15.5 (3.1)	17.0 (3.3)	15.5 (3.2)	16.2 (3.5)
Pageblocks	66.6 (4.3)	57.5 (6.6)	66.7 (5.2)	67.6 (4.0)	59.2 (8.1)	55.9 (6.4)	62.6 (7.6)	59.0 (7.8)
Yeast	36.2 (12.9)	27.2 (8.5)	38.6 (12.1)	37.4 (10.1)	39.9 (6.5)	27.6 (6.8)	39.3 (4.3)	37.9 (4.8)
Wine	11.0 (6.1)	24.7 (2.0)	14.2 (9.3)	19.3 (7.9)	21.5 (3.7)	25.2 (4.5)	18.6 (5.8)	22.4 (6.4)
Letter	75.4 (0.7)	74.9 (0.8)	80.8 (0.5)	81.0 (0.4)	82.9 (0.3)	82.9 (0.3)	82.9 (0.2)	82.9 (0.2)
News20	90.9 (0.1)	91.0 (0.2)	91.1 (0.1)	91.0 (0.1)	90.6 (0.1)	90.6 (0.1)	90.6 (0.2)	90.6 (0.2)
Average	47.4 (3.5)	47.5 (2.6)	49.8 (3.7)	50.4 (3.0)	48.8 (3.6)	48.2 (2.6)	49.1 (3.6)	49.1 (3.8)

Finally, we give here exhaustive tabular results, giving test-time F-measure results obtained by different methods when varying the budget (when meaningful) from 1 to 18 call to the weight classifier learning algorithm to complete the previous graphs.

Table 3: Mean F-Measure over 5 experiments and **limiting the number of iterations/grid steps to 1** (standard deviation between brackets).

Datasets	SVM	$SVM_{I.R.}$	SVM_G	SVM_C	LR	$LR_{I.R.}$	LR_B	LR_G	LR_C
Adult	62.5 (0.2)	64.9 (0.3)	65.0 (0.4)	65.0 (0.4)	63.1 (0.1)	66.0 (0.1)	66.6 (0.1)	66.1 (0.1)	66.1 (0.1)
Abalone10	0.0 (0.0)	30.9 (1.2)	0.0 (0.0)	0.0 (0.0)	0.0 (0.0)	31.9 (1.4)	31.6 (0.6)	24.4 (1.2)	24.4 (1.3)
Satimage	0.0 (0.0)	23.4 (4.3)	0.9 (1.9)	0.0 (0.0)	0.5 (0.9)	24.2 (5.3)	21.4 (4.6)	3.5 (6.9)	3.5 (6.9)
IJCNN	44.5 (0.4)	53.3 (0.4)	61.6 (0.5)	61.6 (0.5)	46.2 (0.3)	51.6 (0.3)	59.2 (0.3)	58.3 (0.3)	58.3 (0.3)
Abalone12	0.0 (0.0)	16.8 (2.7)	0.0 (0.0)	0.0 (0.0)	0.0 (0.0)	18.0 (3.5)	17.7 (3.7)	0.0 (0.0)	0.0 (0.0)
Pageblocks	48.1 (5.8)	39.6 (4.7)	64.4 (2.9)	59.1 (3.8)	48.6 (3.3)	42.4 (5.2)	55.7 (5.7)	55.3 (4.7)	54.5 (4.4)
Yeast	0.0 (0.0)	29.4 (2.9)	12.1 (10.6)	22.9 (15.7)	2.5 (5.0)	29.0 (3.5)	35.4 (15.6)	24.9 (16.0)	24.4 (16.1)
Wine	0.0 (0.0)	15.6 (5.2)	0.0 (0.0)	0.0 (0.0)	0.0 (0.0)	14.6 (3.2)	18.3 (7.2)	5.5 (10.9)	11.6 (10.8)
Letter	75.4 (0.7)	74.9 (0.8)	80.2 (0.3)	80.3 (0.3)	82.9 (0.3)	82.9 (0.3)	74.9 (0.5)	82.6 (0.3)	82.6 (0.3)
News20	90.9 (0.1)	91.0 (0.2)	90.9 (0.2)	90.9 (0.2)	90.6 (0.1)	90.6 (0.1)	89.4 (0.2)	90.6 (0.2)	90.6 (0.2)
Average	32.1 (0.7)	44.0 (2.3)	37.5 (1.7)	38.0 (2.1)	33.4 (1.0)	45.1 (2.3)	47.0 (3.9)	41.1 (4.1)	41.6 (4.0)

Table 4: Mean F-Measure over 5 experiments and **limiting the number of iterations/grid steps to 2** (standard deviation between brackets).

Datasets	SVM	$SVM_{I.R.}$	SVM_G	SVM_C	LR	$LR_{I.R.}$	LR_B	LR_G	LR_C
Adult	62.5 (0.2)	64.9 (0.3)	66.4 (0.2)	66.2 (0.3)	63.1 (0.1)	66.0 (0.1)	66.6 (0.1)	66.6 (0.1)	66.2 (0.1)
Abalone10	0.0 (0.0)	30.9 (1.2)	32.6 (1.4)	30.7 (1.1)	0.0 (0.0)	31.9 (1.4)	31.6 (0.6)	31.9 (1.7)	32.4 (1.9)
Satimage	0.0 (0.0)	23.4 (4.3)	6.1 (12.2)	5.9 (11.8)	0.5 (0.9)	24.2 (5.3)	21.4 (4.6)	6.2 (12.3)	6.1 (12.2)
IJCNN	44.5 (0.4)	53.3 (0.4)	60.7 (0.4)	61.6 (0.5)	46.2 (0.3)	51.6 (0.3)	59.2 (0.3)	56.8 (0.3)	58.3 (0.3)
Abalone12	0.0 (0.0)	16.8 (2.7)	0.0 (0.0)	0.0 (0.0)	0.0 (0.0)	18.0 (3.5)	17.7 (3.7)	2.8 (3.4)	13.3 (3.5)
Pageblocks	48.1 (5.8)	39.6 (4.7)	65.0 (7.6)	63.3 (4.1)	48.6 (3.3)	42.4 (5.2)	55.7 (5.7)	62.7 (7.1)	58.3 (6.8)
Yeast	0.0 (0.0)	29.4 (2.9)	30.9 (17.2)	25.4 (17.5)	2.5 (5.0)	29.0 (3.5)	35.4 (15.6)	27.8 (20.0)	33.0 (18.3)
Wine	0.0 (0.0)	15.6 (5.2)	0.0 (0.0)	11.7 (11.1)	0.0 (0.0)	14.6 (3.2)	18.3 (7.2)	8.7 (11.2)	15.6 (6.7)
Letter	75.4 (0.7)	74.9 (0.8)	80.7 (0.5)	80.4 (0.5)	82.9 (0.3)	82.9 (0.3)	74.9 (0.5)	82.9 (0.2)	82.8 (0.2)
News20	90.9 (0.1)	91.0 (0.2)	90.9 (0.2)	91.0 (0.2)	90.6 (0.1)	90.6 (0.1)	89.4 (0.2)	90.6 (0.2)	90.6 (0.1)
Average	32.1 (0.7)	44.0 (2.3)	43.3 (4.0)	43.6 (4.7)	33.4 (1.0)	45.1 (2.3)	47.0 (3.9)	43.7 (5.6)	45.7 (5.0)

Table 5: Mean F-Measure over 5 experiments and **limiting the number of iterations/grid steps to 3** (standard deviation between brackets).

Datasets	SVM	$SVM_{I.R.}$	SVM_G	SVM_C	LR	$LR_{I.R.}$	LR_B	LR_G	LR_C
Adult	62.5 (0.2)	64.9 (0.3)	66.1 (0.2)	66.2 (0.3)	63.1 (0.1)	66.0 (0.1)	66.6 (0.1)	66.2 (0.1)	66.2 (0.1)
Abalone10	0.0 (0.0)	30.9 (1.2)	30.7 (1.1)	31.0 (1.4)	0.0 (0.0)	31.9 (1.4)	31.6 (0.6)	32.5 (1.5)	31.3 (2.2)
Satimage	0.0 (0.0)	23.4 (4.3)	5.9 (11.8)	20.2 (4.7)	0.5 (0.9)	24.2 (5.3)	21.4 (4.6)	6.1 (12.1)	20.3 (5.1)
IJCNN	44.5 (0.4)	53.3 (0.4)	61.6 (0.5)	61.6 (0.5)	46.2 (0.3)	51.6 (0.3)	59.2 (0.3)	58.3 (0.3)	58.3 (0.3)
Abalone12	0.0 (0.0)	16.8 (2.7)	0.0 (0.0)	16.7 (2.7)	0.0 (0.0)	18.0 (3.5)	17.7 (3.7)	14.2 (3.0)	16.6 (3.4)
Pageblocks	48.1 (5.8)	39.6 (4.7)	65.5 (2.0)	63.3 (4.1)	48.6 (3.3)	42.4 (5.2)	55.7 (5.7)	60.4 (6.4)	58.3 (6.8)
Yeast	0.0 (0.0)	29.4 (2.9)	32.6 (18.3)	37.8 (7.8)	2.5 (5.0)	29.0 (3.5)	35.4 (15.6)	32.1 (11.9)	32.6 (12.0)
Wine	0.0 (0.0)	15.6 (5.2)	11.8 (11.1)	19.5 (5.1)	0.0 (0.0)	14.6 (3.2)	18.3 (7.2)	17.5 (5.8)	20.0 (3.8)
Letter	75.4 (0.7)	74.9 (0.8)	80.5 (0.2)	80.4 (0.5)	82.9 (0.3)	82.9 (0.3)	74.9 (0.5)	82.9 (0.2)	82.9 (0.2)
News20	90.9 (0.1)	91.0 (0.2)	91.0 (0.2)	91.0 (0.2)	90.6 (0.1)	90.6 (0.1)	89.4 (0.2)	90.6 (0.2)	90.6 (0.1)
Average	32.1 (0.7)	44.0 (2.3)	44.6 (4.5)	48.8 (2.7)	33.4 (1.0)	45.1 (2.3)	47.0 (3.9)	46.1 (4.2)	47.7 (3.4)

Table 6: Mean F-Measure over 5 experiments and **limiting the number of iterations/grid steps to 4** (standard deviation between brackets).

Datasets	SVM	$SVM_{I.R.}$	SVM_G	SVM_C	LR	$LR_{I.R.}$	LR_B	LR_G	LR_C
Adult	62.5 (0.2)	64.9 (0.3)	66.0 (0.2)	66.2 (0.3)	63.1 (0.1)	66.0 (0.1)	66.6 (0.1)	66.4 (0.1)	66.2 (0.1)
Abalone10	0.0 (0.0)	30.9 (1.2)	31.0 (1.0)	31.0 (1.4)	0.0 (0.0)	31.9 (1.4)	31.6 (0.6)	30.9 (1.7)	31.3 (2.2)
Satimage	0.0 (0.0)	23.4 (4.3)	16.4 (9.5)	20.6 (5.6)	0.5 (0.9)	24.2 (5.3)	21.4 (4.6)	17.0 (9.8)	20.5 (5.0)
IJCNN	44.5 (0.4)	53.3 (0.4)	61.5 (0.4)	61.1 (0.5)	46.2 (0.3)	51.6 (0.3)	59.2 (0.3)	57.8 (0.4)	58.3 (0.3)
Abalone12	0.0 (0.0)	16.8 (2.7)	16.5 (4.0)	16.9 (4.3)	0.0 (0.0)	18.0 (3.5)	17.7 (3.7)	17.6 (3.0)	17.6 (3.1)
Pageblocks	48.1 (5.8)	39.6 (4.7)	61.0 (6.0)	63.3 (4.1)	48.6 (3.3)	42.4 (5.2)	55.7 (5.7)	62.1 (7.8)	58.4 (6.7)
Yeast	0.0 (0.0)	29.4 (2.9)	35.4 (8.7)	39.0 (7.5)	2.5 (5.0)	29.0 (3.5)	35.4 (15.6)	31.1 (18.0)	32.5 (12.0)
Wine	0.0 (0.0)	15.6 (5.2)	11.5 (7.8)	19.5 (5.1)	0.0 (0.0)	14.6 (3.2)	18.3 (7.2)	17.9 (2.8)	20.0 (3.8)
Letter	75.4 (0.7)	74.9 (0.8)	80.5 (0.3)	80.4 (0.5)	82.9 (0.3)	82.9 (0.3)	74.9 (0.5)	82.9 (0.2)	82.9 (0.3)
News20	90.9 (0.1)	91.0 (0.2)	91.0 (0.1)	91.0 (0.2)	90.6 (0.1)	90.6 (0.1)	89.4 (0.2)	90.6 (0.2)	90.7 (0.1)
Average	32.1 (0.7)	44.0 (2.3)	47.1 (3.8)	48.9 (2.9)	33.4 (1.0)	45.1 (2.3)	47.0 (3.9)	47.4 (4.4)	47.8 (3.4)

Table 7: Mean F-Measure over 5 experiments and **limiting the number of iterations/grid steps to 5** (standard deviation between brackets).

Datasets	SVM	$SVM_{I.R.}$	SVM_G	SVM_C	LR	$LR_{I.R.}$	LR_B	LR_G	LR_C
Adult	62.5 (0.2)	64.9 (0.3)	66.4 (0.3)	66.2 (0.2)	63.1 (0.1)	66.0 (0.1)	66.6 (0.1)	66.6 (0.1)	66.5 (0.1)
Abalone10	0.0 (0.0)	30.9 (1.2)	32.6 (1.4)	31.7 (1.0)	0.0 (0.0)	31.9 (1.4)	31.6 (0.6)	31.3 (0.7)	31.2 (2.3)
Satimage	0.0 (0.0)	23.4 (4.3)	16.4 (9.5)	20.6 (5.6)	0.5 (0.9)	24.2 (5.3)	21.4 (4.6)	17.0 (9.8)	20.5 (5.0)
IJCNN	44.5 (0.4)	53.3 (0.4)	61.4 (0.6)	61.1 (0.5)	46.2 (0.3)	51.6 (0.3)	59.2 (0.3)	58.3 (0.3)	58.3 (0.3)
Abalone12	0.0 (0.0)	16.8 (2.7)	16.5 (4.0)	16.5 (4.0)	0.0 (0.0)	18.0 (3.5)	17.7 (3.7)	17.6 (3.0)	18.1 (2.6)
Pageblocks	48.1 (5.8)	39.6 (4.7)	67.7 (4.0)	62.1 (5.0)	48.6 (3.3)	42.4 (5.2)	55.7 (5.7)	61.8 (7.3)	59.6 (7.3)
Yeast	0.0 (0.0)	29.4 (2.9)	31.8 (10.5)	39.0 (7.5)	2.5 (5.0)	29.0 (3.5)	35.4 (15.6)	30.1 (17.2)	38.8 (8.5)
Wine	0.0 (0.0)	15.6 (5.2)	11.5 (7.8)	20.4 (5.6)	0.0 (0.0)	14.6 (3.2)	18.3 (7.2)	17.9 (2.8)	21.2 (5.1)
Letter	75.4 (0.7)	74.9 (0.8)	80.5 (0.4)	80.4 (0.5)	82.9 (0.3)	82.9 (0.3)	74.9 (0.5)	82.9 (0.2)	82.9 (0.3)
News20	90.9 (0.1)	91.0 (0.2)	91.0 (0.1)	91.0 (0.2)	90.6 (0.1)	90.6 (0.1)	89.4 (0.2)	90.6 (0.2)	90.7 (0.1)
Average	32.1 (0.7)	44.0 (2.3)	47.6 (3.9)	48.9 (3.0)	33.4 (1.0)	45.1 (2.3)	47.0 (3.9)	47.4 (4.2)	48.8 (3.2)

Table 8: Mean F-Measure over 5 experiments and **limiting the number of iterations/grid steps to 6** (standard deviation between brackets).

Datasets	SVM	$SVM_{I.R.}$	SVM_G	SVM_C	LR	$LR_{I.R.}$	LR_B	LR_G	LR_C
Adult	62.5 (0.2)	64.9 (0.3)	66.5 (0.1)	66.4 (0.1)	63.1 (0.1)	66.0 (0.1)	66.6 (0.1)	66.4 (0.2)	66.5 (0.1)
Abalone10	0.0 (0.0)	30.9 (1.2)	30.7 (1.1)	31.7 (1.0)	0.0 (0.0)	31.9 (1.4)	31.6 (0.6)	31.6 (1.0)	31.4 (2.2)
Satimage	0.0 (0.0)	23.4 (4.3)	20.4 (5.3)	20.6 (5.6)	0.5 (0.9)	24.2 (5.3)	21.4 (4.6)	20.1 (4.6)	20.5 (5.0)
IJCNN	44.5 (0.4)	53.3 (0.4)	62.1 (0.5)	61.3 (0.6)	46.2 (0.3)	51.6 (0.3)	59.2 (0.3)	58.0 (0.4)	58.1 (0.3)
Abalone12	0.0 (0.0)	16.8 (2.7)	16.9 (2.9)	18.2 (3.3)	0.0 (0.0)	18.0 (3.5)	17.7 (3.7)	15.5 (6.2)	17.7 (3.4)
Pageblocks	48.1 (5.8)	39.6 (4.7)	64.8 (3.1)	64.2 (4.6)	48.6 (3.3)	42.4 (5.2)	55.7 (5.7)	60.6 (9.1)	59.5 (7.4)
Yeast	0.0 (0.0)	29.4 (2.9)	32.0 (10.4)	39.0 (7.5)	2.5 (5.0)	29.0 (3.5)	35.4 (15.6)	33.0 (18.8)	38.8 (8.5)
Wine	0.0 (0.0)	15.6 (5.2)	19.4 (5.3)	19.0 (7.0)	0.0 (0.0)	14.6 (3.2)	18.3 (7.2)	19.6 (5.1)	21.1 (5.2)
Letter	75.4 (0.7)	74.9 (0.8)	80.8 (0.5)	80.5 (0.4)	82.9 (0.3)	82.9 (0.3)	74.9 (0.5)	82.9 (0.2)	82.9 (0.3)
News20	90.9 (0.1)	91.0 (0.2)	91.0 (0.2)	91.0 (0.2)	90.6 (0.1)	90.6 (0.1)	89.4 (0.2)	90.6 (0.1)	90.7 (0.1)
Average	32.1 (0.7)	44.0 (2.3)	48.5 (2.9)	49.2 (3.0)	33.4 (1.0)	45.1 (2.3)	47.0 (3.9)	47.8 (4.6)	48.7 (3.2)

Table 9: Mean F-Measure over 5 experiments and **limiting the number of iterations/grid steps to 7** (standard deviation between brackets).

Datasets	SVM	$SVM_{I.R.}$	SVM_G	SVM_C	LR	$LR_{I.R.}$	LR_B	LR_G	LR_C
Adult	62.5 (0.2)	64.9 (0.3)	66.2 (0.1)	66.4 (0.1)	63.1 (0.1)	66.0 (0.1)	66.6 (0.1)	66.4 (0.1)	66.5 (0.1)
Abalone10	0.0 (0.0)	30.9 (1.2)	31.0 (1.0)	32.5 (1.0)	0.0 (0.0)	31.9 (1.4)	31.6 (0.6)	32.2 (0.6)	31.4 (2.2)
Satimage	0.0 (0.0)	23.4 (4.3)	20.2 (4.7)	20.6 (5.6)	0.5 (0.9)	24.2 (5.3)	21.4 (4.6)	20.3 (5.0)	20.5 (5.0)
IJCNN	44.5 (0.4)	53.3 (0.4)	61.5 (0.4)	61.5 (0.5)	46.2 (0.3)	51.6 (0.3)	59.2 (0.3)	58.3 (0.3)	58.1 (0.3)
Abalone12	0.0 (0.0)	16.8 (2.7)	16.9 (2.9)	18.3 (3.3)	0.0 (0.0)	18.0 (3.5)	17.7 (3.7)	17.5 (3.4)	17.7 (3.4)
Pageblocks	48.1 (5.8)	39.6 (4.7)	65.7 (2.6)	62.8 (3.9)	48.6 (3.3)	42.4 (5.2)	55.7 (5.7)	61.3 (9.9)	59.9 (7.0)
Yeast	0.0 (0.0)	29.4 (2.9)	38.8 (7.0)	39.0 (7.5)	2.5 (5.0)	29.0 (3.5)	35.4 (15.6)	32.7 (11.8)	38.9 (8.6)
Wine	0.0 (0.0)	15.6 (5.2)	19.5 (6.2)	19.0 (7.0)	0.0 (0.0)	14.6 (3.2)	18.3 (7.2)	18.7 (4.5)	21.1 (5.2)
Letter	75.4 (0.7)	74.9 (0.8)	80.6 (0.1)	80.5 (0.4)	82.9 (0.3)	82.9 (0.3)	74.9 (0.5)	82.9 (0.2)	82.9 (0.3)
News20	90.9 (0.1)	91.0 (0.2)	91.1 (0.1)	91.0 (0.2)	90.6 (0.1)	90.6 (0.1)	89.4 (0.2)	90.6 (0.1)	90.7 (0.1)
Average	32.1 (0.7)	44.0 (2.3)	49.2 (2.5)	49.2 (3.0)	33.4 (1.0)	45.1 (2.3)	47.0 (3.9)	48.1 (3.6)	48.8 (3.2)

Table 10: Mean F-Measure over 5 experiments and **limiting the number of iterations/grid steps to 8** (standard deviation between brackets).

Datasets	SVM	$SVM_{I.R.}$	SVM_G	SVM_C	LR	$LR_{I.R.}$	LR_B	LR_G	LR_C
Adult	62.5 (0.2)	64.9 (0.3)	66.4 (0.1)	66.5 (0.1)	63.1 (0.1)	66.0 (0.1)	66.6 (0.1)	66.5 (0.1)	66.5 (0.1)
Abalone10	0.0 (0.0)	30.9 (1.2)	32.6 (1.4)	32.6 (1.0)	0.0 (0.0)	31.9 (1.4)	31.6 (0.6)	32.1 (0.8)	31.4 (2.2)
Satimage	0.0 (0.0)	23.4 (4.3)	20.2 (4.7)	20.6 (5.6)	0.5 (0.9)	24.2 (5.3)	21.4 (4.6)	20.3 (5.0)	20.5 (5.0)
IJCNN	44.5 (0.4)	53.3 (0.4)	61.9 (0.7)	61.5 (0.5)	46.2 (0.3)	51.6 (0.3)	59.2 (0.3)	58.0 (0.4)	58.1 (0.3)
Abalone12	0.0 (0.0)	16.8 (2.7)	16.9 (2.9)	18.3 (3.3)	0.0 (0.0)	18.0 (3.5)	17.7 (3.7)	17.5 (3.4)	18.1 (3.7)
Pageblocks	48.1 (5.8)	39.6 (4.7)	65.8 (4.3)	62.8 (3.9)	48.6 (3.3)	42.4 (5.2)	55.7 (5.7)	60.0 (8.8)	59.4 (7.5)
Yeast	0.0 (0.0)	29.4 (2.9)	33.3 (12.2)	39.0 (7.5)	2.5 (5.0)	29.0 (3.5)	35.4 (15.6)	39.4 (8.5)	38.9 (8.6)
Wine	0.0 (0.0)	15.6 (5.2)	19.5 (6.2)	22.4 (6.1)	0.0 (0.0)	14.6 (3.2)	18.3 (7.2)	18.7 (4.5)	21.1 (5.2)
Letter	75.4 (0.7)	74.9 (0.8)	80.6 (0.4)	80.5 (0.4)	82.9 (0.3)	82.9 (0.3)	74.9 (0.5)	82.9 (0.2)	82.9 (0.3)
News20	90.9 (0.1)	91.0 (0.2)	91.0 (0.1)	91.0 (0.2)	90.6 (0.1)	90.6 (0.1)	89.4 (0.2)	90.6 (0.1)	90.6 (0.1)
Average	32.1 (0.7)	44.0 (2.3)	48.8 (3.3)	49.5 (2.9)	33.4 (1.0)	45.1 (2.3)	47.0 (3.9)	48.6 (3.2)	48.8 (3.3)

Table 11: Mean F-Measure over 5 experiments and **limiting the number of iterations/grid steps to 9** (standard deviation between brackets).

Datasets	SVM	$SVM_{I.R.}$	SVM_G	SVM_C	LR	$LR_{I.R.}$	LR_B	LR_G	LR_C
Adult	62.5 (0.2)	64.9 (0.3)	66.4 (0.1)	66.5 (0.1)	63.1 (0.1)	66.0 (0.1)	66.6 (0.1)	66.4 (0.1)	66.5 (0.1)
Abalone10	0.0 (0.0)	30.9 (1.2)	31.0 (1.0)	32.2 (0.8)	0.0 (0.0)	31.9 (1.4)	31.6 (0.6)	31.5 (0.4)	31.4 (2.2)
Satimage	0.0 (0.0)	23.4 (4.3)	20.4 (5.3)	20.6 (5.6)	0.5 (0.9)	24.2 (5.3)	21.4 (4.6)	20.8 (4.9)	20.5 (5.0)
IJCNN	44.5 (0.4)	53.3 (0.4)	61.5 (0.4)	61.5 (0.5)	46.2 (0.3)	51.6 (0.3)	59.2 (0.3)	58.3 (0.3)	58.1 (0.3)
Abalone12	0.0 (0.0)	16.8 (2.7)	16.7 (4.1)	18.3 (3.3)	0.0 (0.0)	18.0 (3.5)	17.7 (3.7)	15.1 (5.9)	18.0 (3.6)
Pageblocks	48.1 (5.8)	39.6 (4.7)	65.4 (2.3)	62.8 (3.9)	48.6 (3.3)	42.4 (5.2)	55.7 (5.7)	62.7 (8.3)	59.4 (7.5)
Yeast	0.0 (0.0)	29.4 (2.9)	38.3 (3.8)	39.0 (7.5)	2.5 (5.0)	29.0 (3.5)	35.4 (15.6)	38.9 (10.9)	38.9 (8.6)
Wine	0.0 (0.0)	15.6 (5.2)	15.5 (6.0)	22.7 (6.0)	0.0 (0.0)	14.6 (3.2)	18.3 (7.2)	20.7 (6.0)	21.1 (5.2)
Letter	75.4 (0.7)	74.9 (0.8)	80.8 (0.5)	80.5 (0.5)	82.9 (0.3)	82.9 (0.3)	74.9 (0.5)	82.9 (0.2)	82.9 (0.3)
News20	90.9 (0.1)	91.0 (0.2)	91.1 (0.1)	91.0 (0.2)	90.6 (0.1)	90.6 (0.1)	89.4 (0.2)	90.6 (0.2)	90.6 (0.1)
Average	32.1 (0.7)	44.0 (2.3)	48.7 (2.4)	49.5 (2.8)	33.4 (1.0)	45.1 (2.3)	47.0 (3.9)	48.8 (3.7)	48.7 (3.3)

Table 12: Mean F-Measure over 5 experiments and **limiting the number of iterations/grid steps to 10** (standard deviation between brackets).

Datasets	SVM	$SVM_{I.R.}$	SVM_G	SVM_C	LR	$LR_{I.R.}$	LR_B	LR_G	LR_C
Adult	62.5 (0.2)	64.9 (0.3)	66.5 (0.1)	66.4 (0.1)	63.1 (0.1)	66.0 (0.1)	66.6 (0.1)	66.5 (0.1)	66.5 (0.1)
Abalone10	0.0 (0.0)	30.9 (1.2)	32.6 (1.4)	32.2 (0.8)	0.0 (0.0)	31.9 (1.4)	31.6 (0.6)	31.8 (1.0)	31.1 (2.0)
Satimage	0.0 (0.0)	23.4 (4.3)	20.4 (5.3)	20.6 (5.6)	0.5 (0.9)	24.2 (5.3)	21.4 (4.6)	20.8 (4.9)	20.5 (5.0)
IJCNN	44.5 (0.4)	53.3 (0.4)	61.9 (0.7)	61.5 (0.5)	46.2 (0.3)	51.6 (0.3)	59.2 (0.3)	58.0 (0.4)	58.1 (0.3)
Abalone12	0.0 (0.0)	16.8 (2.7)	16.7 (4.1)	18.3 (3.3)	0.0 (0.0)	18.0 (3.5)	17.7 (3.7)	15.1 (5.9)	17.8 (3.4)
Pageblocks	48.1 (5.8)	39.6 (4.7)	65.6 (4.1)	62.8 (3.9)	48.6 (3.3)	42.4 (5.2)	55.7 (5.7)	61.3 (7.3)	59.4 (7.5)
Yeast	0.0 (0.0)	29.4 (2.9)	32.5 (10.4)	39.0 (7.5)	2.5 (5.0)	29.0 (3.5)	35.4 (15.6)	38.9 (10.9)	39.5 (9.3)
Wine	0.0 (0.0)	15.6 (5.2)	15.5 (6.0)	22.7 (6.0)	0.0 (0.0)	14.6 (3.2)	18.3 (7.2)	20.7 (6.0)	21.1 (5.2)
Letter	75.4 (0.7)	74.9 (0.8)	80.8 (0.5)	80.7 (0.4)	82.9 (0.3)	82.9 (0.3)	74.9 (0.5)	82.9 (0.2)	82.9 (0.3)
News20	90.9 (0.1)	91.0 (0.2)	91.0 (0.1)	91.0 (0.2)	90.6 (0.1)	90.6 (0.1)	89.4 (0.2)	90.6 (0.2)	90.6 (0.1)
Average	32.1 (0.7)	44.0 (2.3)	48.4 (3.3)	49.5 (2.8)	33.4 (1.0)	45.1 (2.3)	47.0 (3.9)	48.7 (3.7)	48.8 (3.3)

Table 13: Mean F-Measure over 5 experiments and **limiting the number of iterations/grid steps to 11** (standard deviation between brackets).

Datasets	SVM	$SVM_{I.R.}$	SVM_G	SVM_C	LR	$LR_{I.R.}$	LR_B	LR_G	LR_C
Adult	62.5 (0.2)	64.9 (0.3)	66.4 (0.1)	66.5 (0.1)	63.1 (0.1)	66.0 (0.1)	66.6 (0.1)	66.5 (0.1)	66.5 (0.1)
Abalone10	0.0 (0.0)	30.9 (1.2)	32.4 (1.3)	32.2 (0.8)	0.0 (0.0)	31.9 (1.4)	31.6 (0.6)	31.9 (0.7)	30.9 (1.9)
Satimage	0.0 (0.0)	23.4 (4.3)	20.2 (4.7)	20.6 (5.6)	0.5 (0.9)	24.2 (5.3)	21.4 (4.6)	20.7 (4.8)	20.5 (5.0)
IJCNN	44.5 (0.4)	53.3 (0.4)	61.4 (0.5)	61.8 (0.5)	46.2 (0.3)	51.6 (0.3)	59.2 (0.3)	58.3 (0.3)	58.1 (0.4)
Abalone12	0.0 (0.0)	16.8 (2.7)	16.7 (4.1)	18.3 (3.3)	0.0 (0.0)	18.0 (3.5)	17.7 (3.7)	17.2 (3.1)	17.8 (3.4)
Pageblocks	48.1 (5.8)	39.6 (4.7)	66.4 (3.5)	62.8 (3.9)	48.6 (3.3)	42.4 (5.2)	55.7 (5.7)	62.6 (8.0)	59.4 (7.5)
Yeast	0.0 (0.0)	29.4 (2.9)	38.4 (7.1)	39.0 (7.5)	2.5 (5.0)	29.0 (3.5)	35.4 (15.6)	38.7 (8.1)	39.5 (9.3)
Wine	0.0 (0.0)	15.6 (5.2)	16.4 (5.9)	22.7 (6.0)	0.0 (0.0)	14.6 (3.2)	18.3 (7.2)	20.5 (6.0)	21.1 (5.2)
Letter	75.4 (0.7)	74.9 (0.8)	80.7 (0.3)	80.9 (0.4)	82.9 (0.3)	82.9 (0.3)	74.9 (0.5)	82.9 (0.2)	82.9 (0.3)
News20	90.9 (0.1)	91.0 (0.2)	91.0 (0.2)	91.0 (0.2)	90.6 (0.1)	90.6 (0.1)	89.4 (0.2)	90.6 (0.2)	90.6 (0.1)
Average	32.1 (0.7)	44.0 (2.3)	49.0 (2.8)	49.6 (2.8)	33.4 (1.0)	45.1 (2.3)	47.0 (3.9)	49.0 (3.1)	48.7 (3.3)

Table 14: Mean F-Measure over 5 experiments and **limiting the number of iterations/grid steps to 12** (standard deviation between brackets).

Datasets	SVM	$SVM_{I.R.}$	SVM_G	SVM_C	LR	$LR_{I.R.}$	LR_B	LR_G	LR_C
Adult	62.5 (0.2)	64.9 (0.3)	66.4 (0.1)	66.5 (0.1)	63.1 (0.1)	66.0 (0.1)	66.6 (0.1)	66.4 (0.1)	66.5 (0.1)
Abalone10	0.0 (0.0)	30.9 (1.2)	31.0 (1.0)	32.2 (0.8)	0.0 (0.0)	31.9 (1.4)	31.6 (0.6)	32.0 (0.7)	30.9 (1.9)
Satimage	0.0 (0.0)	23.4 (4.3)	20.4 (5.3)	20.6 (5.6)	0.5 (0.9)	24.2 (5.3)	21.4 (4.6)	20.3 (5.0)	20.5 (5.0)
IJCNN	44.5 (0.4)	53.3 (0.4)	61.8 (0.4)	61.6 (0.6)	46.2 (0.3)	51.6 (0.3)	59.2 (0.3)	58.0 (0.4)	58.2 (0.3)
Abalone12	0.0 (0.0)	16.8 (2.7)	16.9 (2.9)	18.3 (3.3)	0.0 (0.0)	18.0 (3.5)	17.7 (3.7)	17.5 (3.4)	17.8 (3.4)
Pageblocks	48.1 (5.8)	39.6 (4.7)	64.7 (3.2)	62.8 (3.9)	48.6 (3.3)	42.4 (5.2)	55.7 (5.7)	61.5 (10.0)	59.4 (7.5)
Yeast	0.0 (0.0)	29.4 (2.9)	38.1 (7.6)	39.0 (7.5)	2.5 (5.0)	29.0 (3.5)	35.4 (15.6)	39.1 (10.1)	39.5 (9.3)
Wine	0.0 (0.0)	15.6 (5.2)	20.0 (6.4)	22.7 (6.0)	0.0 (0.0)	14.6 (3.2)	18.3 (7.2)	18.7 (4.5)	21.1 (5.2)
Letter	75.4 (0.7)	74.9 (0.8)	80.8 (0.5)	80.9 (0.4)	82.9 (0.3)	82.9 (0.3)	74.9 (0.5)	82.9 (0.3)	82.9 (0.3)
News20	90.9 (0.1)	91.0 (0.2)	91.0 (0.1)	91.0 (0.2)	90.6 (0.1)	90.6 (0.1)	89.4 (0.2)	90.6 (0.1)	90.6 (0.2)
Average	32.1 (0.7)	44.0 (2.3)	49.1 (2.7)	49.6 (2.8)	33.4 (1.0)	45.1 (2.3)	47.0 (3.9)	48.7 (3.5)	48.7 (3.3)

Table 15: Mean F-Measure over 5 experiments and **limiting the number of iterations/grid steps to 13** (standard deviation between brackets).

Datasets	SVM	$SVM_{I.R.}$	SVM_G	SVM_C	LR	$LR_{I.R.}$	LR_B	LR_G	LR_C
Adult	62.5 (0.2)	64.9 (0.3)	66.4 (0.1)	66.5 (0.1)	63.1 (0.1)	66.0 (0.1)	66.6 (0.1)	66.5 (0.1)	66.5 (0.1)
Abalone10	0.0 (0.0)	30.9 (1.2)	32.6 (1.4)	32.2 (0.8)	0.0 (0.0)	31.9 (1.4)	31.6 (0.6)	32.3 (1.1)	30.9 (1.9)
Satimage	0.0 (0.0)	23.4 (4.3)	20.4 (5.3)	20.6 (5.6)	0.5 (0.9)	24.2 (5.3)	21.4 (4.6)	20.3 (5.0)	20.5 (5.0)
IJCNN	44.5 (0.4)	53.3 (0.4)	61.9 (0.7)	61.6 (0.6)	46.2 (0.3)	51.6 (0.3)	59.2 (0.3)	58.2 (0.2)	58.2 (0.3)
Abalone12	0.0 (0.0)	16.8 (2.7)	16.9 (2.9)	18.3 (3.3)	0.0 (0.0)	18.0 (3.5)	17.7 (3.7)	17.5 (3.4)	17.8 (3.4)
Pageblocks	48.1 (5.8)	39.6 (4.7)	66.6 (3.1)	62.8 (3.9)	48.6 (3.3)	42.4 (5.2)	55.7 (5.7)	60.2 (9.0)	59.4 (7.5)
Yeast	0.0 (0.0)	29.4 (2.9)	33.3 (12.2)	39.0 (7.5)	2.5 (5.0)	29.0 (3.5)	35.4 (15.6)	39.1 (10.1)	39.5 (9.3)
Wine	0.0 (0.0)	15.6 (5.2)	20.0 (6.4)	22.7 (6.0)	0.0 (0.0)	14.6 (3.2)	18.3 (7.2)	18.7 (4.5)	21.1 (5.2)
Letter	75.4 (0.7)	74.9 (0.8)	80.8 (0.5)	80.9 (0.4)	82.9 (0.3)	82.9 (0.3)	74.9 (0.5)	82.9 (0.3)	82.9 (0.3)
News20	90.9 (0.1)	91.0 (0.2)	91.0 (0.1)	91.0 (0.2)	90.6 (0.1)	90.6 (0.1)	89.4 (0.2)	90.6 (0.1)	90.6 (0.2)
Average	32.1 (0.7)	44.0 (2.3)	49.0 (3.3)	49.6 (2.8)	33.4 (1.0)	45.1 (2.3)	47.0 (3.9)	48.6 (3.4)	48.7 (3.3)

Table 16: Mean F-Measure over 5 experiments and **limiting the number of iterations/grid steps to 14** (standard deviation between brackets).

Datasets	SVM	$SVM_{I.R.}$	SVM_G	SVM_C	LR	$LR_{I.R.}$	LR_B	LR_G	LR_C
Adult	62.5 (0.2)	64.9 (0.3)	66.5 (0.1)	66.5 (0.1)	63.1 (0.1)	66.0 (0.1)	66.6 (0.1)	66.5 (0.1)	66.5 (0.1)
Abalone10	0.0 (0.0)	30.9 (1.2)	32.4 (1.3)	32.2 (0.8)	0.0 (0.0)	31.9 (1.4)	31.6 (0.6)	31.4 (0.5)	30.9 (1.9)
Satimage	0.0 (0.0)	23.4 (4.3)	20.4 (5.3)	20.6 (5.6)	0.5 (0.9)	24.2 (5.3)	21.4 (4.6)	20.8 (4.9)	20.5 (5.0)
IJCNN	44.5 (0.4)	53.3 (0.4)	61.6 (0.6)	61.6 (0.6)	46.2 (0.3)	51.6 (0.3)	59.2 (0.3)	58.0 (0.4)	58.2 (0.3)
Abalone12	0.0 (0.0)	16.8 (2.7)	16.8 (4.2)	18.3 (3.3)	0.0 (0.0)	18.0 (3.5)	17.7 (3.7)	15.1 (5.9)	17.8 (3.4)
Pageblocks	48.1 (5.8)	39.6 (4.7)	65.5 (4.2)	62.8 (3.9)	48.6 (3.3)	42.4 (5.2)	55.7 (5.7)	62.8 (8.2)	59.4 (7.5)
Yeast	0.0 (0.0)	29.4 (2.9)	38.0 (4.4)	39.0 (7.5)	2.5 (5.0)	29.0 (3.5)	35.4 (15.6)	38.2 (11.2)	39.5 (9.3)
Wine	0.0 (0.0)	15.6 (5.2)	19.1 (6.9)	22.7 (6.0)	0.0 (0.0)	14.6 (3.2)	18.3 (7.2)	18.9 (4.6)	21.1 (5.2)
Letter	75.4 (0.7)	74.9 (0.8)	80.8 (0.5)	80.9 (0.4)	82.9 (0.3)	82.9 (0.3)	74.9 (0.5)	82.9 (0.2)	82.9 (0.3)
News20	90.9 (0.1)	91.0 (0.2)	91.1 (0.1)	91.0 (0.2)	90.6 (0.1)	90.6 (0.1)	89.4 (0.2)	90.6 (0.2)	90.6 (0.2)
Average	32.1 (0.7)	44.0 (2.3)	49.2 (2.8)	49.6 (2.8)	33.4 (1.0)	45.1 (2.3)	47.0 (3.9)	48.5 (3.6)	48.7 (3.3)

Table 17: Mean F-Measure over 5 experiments and **limiting the number of iterations/grid steps to 15** (standard deviation between brackets).

Datasets	SVM	$SVM_{I.R.}$	SVM_G	SVM_C	LR	$LR_{I.R.}$	LR_B	LR_G	LR_C
Adult	62.5 (0.2)	64.9 (0.3)	66.4 (0.1)	66.5 (0.1)	63.1 (0.1)	66.0 (0.1)	66.6 (0.1)	66.4 (0.1)	66.5 (0.1)
Abalone10	0.0 (0.0)	30.9 (1.2)	31.0 (1.0)	32.2 (0.8)	0.0 (0.0)	31.9 (1.4)	31.6 (0.6)	31.9 (0.5)	30.9 (1.9)
Satimage	0.0 (0.0)	23.4 (4.3)	20.4 (5.3)	20.6 (5.6)	0.5 (0.9)	24.2 (5.3)	21.4 (4.6)	20.7 (4.8)	20.5 (5.0)
IJCNN	44.5 (0.4)	53.3 (0.4)	61.8 (0.4)	61.6 (0.6)	46.2 (0.3)	51.6 (0.3)	59.2 (0.3)	58.2 (0.2)	58.2 (0.3)
Abalone12	0.0 (0.0)	16.8 (2.7)	16.8 (4.2)	18.3 (3.3)	0.0 (0.0)	18.0 (3.5)	17.7 (3.7)	17.2 (3.1)	18.4 (2.3)
Pageblocks	48.1 (5.8)	39.6 (4.7)	65.7 (2.1)	62.8 (3.9)	48.6 (3.3)	42.4 (5.2)	55.7 (5.7)	62.7 (8.3)	59.4 (7.5)
Yeast	0.0 (0.0)	29.4 (2.9)	39.0 (6.8)	39.0 (7.5)	2.5 (5.0)	29.0 (3.5)	35.4 (15.6)	39.1 (10.1)	39.5 (9.3)
Wine	0.0 (0.0)	15.6 (5.2)	20.0 (6.4)	22.7 (6.0)	0.0 (0.0)	14.6 (3.2)	18.3 (7.2)	18.7 (4.5)	21.1 (5.2)
Letter	75.4 (0.7)	74.9 (0.8)	80.8 (0.5)	80.9 (0.4)	82.9 (0.3)	82.9 (0.3)	74.9 (0.5)	82.9 (0.2)	82.9 (0.3)
News20	90.9 (0.1)	91.0 (0.2)	91.1 (0.1)	91.0 (0.1)	90.6 (0.1)	90.6 (0.1)	89.4 (0.2)	90.6 (0.2)	90.6 (0.2)
Average	32.1 (0.7)	44.0 (2.3)	49.3 (2.7)	49.6 (2.8)	33.4 (1.0)	45.1 (2.3)	47.0 (3.9)	48.8 (3.2)	48.8 (3.2)

Table 18: Mean F-Measure over 5 experiments and **limiting the number of iterations/grid steps to 16** (standard deviation between brackets).

Datasets	SVM	$SVM_{I.R.}$	SVM_G	SVM_C	LR	$LR_{I.R.}$	LR_B	LR_G	LR_C
Adult	62.5 (0.2)	64.9 (0.3)	66.4 (0.1)	66.5 (0.1)	63.1 (0.1)	66.0 (0.1)	66.6 (0.1)	66.5 (0.1)	66.5 (0.1)
Abalone10	0.0 (0.0)	30.9 (1.2)	32.4 (1.3)	32.2 (0.8)	0.0 (0.0)	31.9 (1.4)	31.6 (0.6)	31.7 (0.7)	30.9 (1.9)
Satimage	0.0 (0.0)	23.4 (4.3)	20.4 (5.3)	20.6 (5.6)	0.5 (0.9)	24.2 (5.3)	21.4 (4.6)	20.7 (4.8)	20.5 (5.0)
IJCNN	44.5 (0.4)	53.3 (0.4)	61.6 (0.6)	61.6 (0.6)	46.2 (0.3)	51.6 (0.3)	59.2 (0.3)	58.0 (0.4)	58.2 (0.3)
Abalone12	0.0 (0.0)	16.8 (2.7)	16.8 (4.2)	18.3 (3.3)	0.0 (0.0)	18.0 (3.5)	17.7 (3.7)	17.2 (3.1)	18.4 (2.3)
Pageblocks	48.1 (5.8)	39.6 (4.7)	65.5 (4.2)	62.8 (3.9)	48.6 (3.3)	42.4 (5.2)	55.7 (5.7)	62.8 (8.2)	59.4 (7.5)
Yeast	0.0 (0.0)	29.4 (2.9)	38.6 (7.1)	39.0 (7.5)	2.5 (5.0)	29.0 (3.5)	35.4 (15.6)	39.1 (10.1)	39.5 (9.3)
Wine	0.0 (0.0)	15.6 (5.2)	20.0 (6.4)	22.7 (6.0)	0.0 (0.0)	14.6 (3.2)	18.3 (7.2)	18.7 (4.5)	21.1 (5.2)
Letter	75.4 (0.7)	74.9 (0.8)	80.8 (0.5)	81.0 (0.4)	82.9 (0.3)	82.9 (0.3)	74.9 (0.5)	82.9 (0.2)	82.9 (0.3)
News20	90.9 (0.1)	91.0 (0.2)	91.1 (0.1)	91.0 (0.1)	90.6 (0.1)	90.6 (0.1)	89.4 (0.2)	90.6 (0.2)	90.6 (0.2)
Average	32.1 (0.7)	44.0 (2.3)	49.4 (3.0)	49.6 (2.8)	33.4 (1.0)	45.1 (2.3)	47.0 (3.9)	48.8 (3.2)	48.8 (3.2)

Table 19: Mean F-Measure over 5 experiments and **limiting the number of iterations/grid steps to 17** (standard deviation between brackets).

Datasets	SVM	$SVM_{I.R.}$	SVM_G	SVM_C	LR	$LR_{I.R.}$	LR_B	LR_G	LR_C
Adult	62.5 (0.2)	64.9 (0.3)	66.4 (0.1)	66.5 (0.1)	63.1 (0.1)	66.0 (0.1)	66.6 (0.1)	66.5 (0.1)	66.5 (0.1)
Abalone10	0.0 (0.0)	30.9 (1.2)	32.4 (1.3)	32.2 (0.8)	0.0 (0.0)	31.9 (1.4)	31.6 (0.6)	31.7 (0.7)	30.9 (1.9)
Satimage	0.0 (0.0)	23.4 (4.3)	20.4 (5.3)	20.6 (5.6)	0.5 (0.9)	24.2 (5.3)	21.4 (4.6)	20.7 (4.8)	20.5 (5.0)
IJCNN	44.5 (0.4)	53.3 (0.4)	61.6 (0.6)	61.6 (0.6)	46.2 (0.3)	51.6 (0.3)	59.2 (0.3)	58.2 (0.2)	58.2 (0.3)
Abalone12	0.0 (0.0)	16.8 (2.7)	16.8 (4.2)	18.3 (3.3)	0.0 (0.0)	18.0 (3.5)	17.7 (3.7)	17.2 (3.1)	18.4 (2.3)
Pageblocks	48.1 (5.8)	39.6 (4.7)	66.4 (3.2)	62.8 (3.9)	48.6 (3.3)	42.4 (5.2)	55.7 (5.7)	62.8 (8.2)	59.4 (7.5)
Yeast	0.0 (0.0)	29.4 (2.9)	38.6 (7.1)	39.0 (7.5)	2.5 (5.0)	29.0 (3.5)	35.4 (15.6)	39.1 (10.1)	39.5 (9.3)
Wine	0.0 (0.0)	15.6 (5.2)	20.0 (6.4)	22.7 (6.0)	0.0 (0.0)	14.6 (3.2)	18.3 (7.2)	18.7 (4.5)	21.1 (5.2)
Letter	75.4 (0.7)	74.9 (0.8)	80.8 (0.5)	81.0 (0.4)	82.9 (0.3)	82.9 (0.3)	74.9 (0.5)	82.9 (0.2)	82.9 (0.3)
News20	90.9 (0.1)	91.0 (0.2)	91.1 (0.1)	91.0 (0.1)	90.6 (0.1)	90.6 (0.1)	89.4 (0.2)	90.6 (0.2)	90.6 (0.2)
Average	32.1 (0.7)	44.0 (2.3)	49.5 (2.9)	49.6 (2.8)	33.4 (1.0)	45.1 (2.3)	47.0 (3.9)	48.8 (3.2)	48.8 (3.2)

Table 20: Mean F-Measure over 5 experiments and **limiting the number of iterations/grid steps to 18** (standard deviation between brackets).

Datasets	SVM	$SVM_{I.R.}$	SVM_G	SVM_C	LR	$LR_{I.R.}$	LR_B	LR_G	LR_C
Adult	62.5 (0.2)	64.9 (0.3)	66.4 (0.1)	66.5 (0.1)	63.1 (0.1)	66.0 (0.1)	66.6 (0.1)	66.5 (0.1)	66.5 (0.1)
Abalone10	0.0 (0.0)	30.9 (1.2)	32.4 (1.3)	32.2 (0.8)	0.0 (0.0)	31.9 (1.4)	31.6 (0.6)	31.7 (0.7)	30.9 (1.9)
Satimage	0.0 (0.0)	23.4 (4.3)	20.4 (5.3)	20.6 (5.6)	0.5 (0.9)	24.2 (5.3)	21.4 (4.6)	20.7 (4.8)	20.5 (5.0)
IJCNN	44.5 (0.4)	53.3 (0.4)	61.6 (0.6)	61.6 (0.6)	46.2 (0.3)	51.6 (0.3)	59.2 (0.3)	58.2 (0.2)	58.2 (0.3)
Abalone12	0.0 (0.0)	16.8 (2.7)	16.8 (4.2)	18.3 (3.3)	0.0 (0.0)	18.0 (3.5)	17.7 (3.7)	17.2 (3.1)	18.4 (2.3)
Pageblocks	48.1 (5.8)	39.6 (4.7)	66.4 (3.2)	62.8 (3.9)	48.6 (3.3)	42.4 (5.2)	55.7 (5.7)	62.8 (8.2)	59.4 (7.5)
Yeast	0.0 (0.0)	29.4 (2.9)	38.6 (7.1)	39.0 (7.5)	2.5 (5.0)	29.0 (3.5)	35.4 (15.6)	39.1 (10.1)	39.5 (9.3)
Wine	0.0 (0.0)	15.6 (5.2)	20.0 (6.4)	22.7 (6.0)	0.0 (0.0)	14.6 (3.2)	18.3 (7.2)	18.7 (4.5)	21.1 (5.2)
Letter	75.4 (0.7)	74.9 (0.8)	80.8 (0.5)	81.0 (0.4)	82.9 (0.3)	82.9 (0.3)	74.9 (0.5)	82.9 (0.2)	82.9 (0.3)
News20	90.9 (0.1)	91.0 (0.2)	91.1 (0.1)	91.0 (0.1)	90.6 (0.1)	90.6 (0.1)	89.4 (0.2)	90.6 (0.2)	90.6 (0.2)
Average	32.1 (0.7)	44.0 (2.3)	49.5 (2.9)	49.6 (2.8)	33.4 (1.0)	45.1 (2.3)	47.0 (3.9)	48.8 (3.2)	48.8 (3.2)

References

Koyejo, O. O., Natarajan, N., Ravikumar, P. K., and Dhillon, I. S. (2014). Consistent binary classification with generalized performance metrics. In Ghahramani, Z., Welling, M., Cortes, C., Lawrence, N. D., and Weinberger, K. Q., editors, *Advances in Neural Information Processing*

Systems 27, pages 2744–2752. Curran Associates, Inc.

Narasimhan, H., Ramaswamy, H., Saha, A., and Agarwal, S. (2015). Consistent multiclass algorithms for complex performance measures. In Bach, F. and Blei, D., editors, *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pages 2398–2407, Lille, France. PMLR.

Parambath, S. P., Usunier, N., and Grandvalet, Y. (2014). Optimizing f-measures by cost-sensitive classification. In *NIPS*, pages 2123–2131.