Une borne transductive multiclasses pour le classifieur de vote majoritaire

Vasilii Feofanov, Emilie Devijver and Massih-Reza Amini

Univ. Grenoble Alpes, CNRS, Grenoble INP, LIG, 38000 Grenoble, France

LIG

UNIVERSITÉ Grenoble Alpes

Objectives

- Semi-supervised learning (SSL) has been well-investigated in the binary classification framework. But there is still a large avenue for theoretical studies for both the binary and the multiclass case.
- In the multiclass framework, there are just few classification methods by now.

In this work we propose:

1. An extension of the self-learning algorithm [1] for the multiclass classification,
2. A transductive bound of the Bayes risk in the multiclass framework.

Framework

We consider the following framework:

- An input $\mathcal{X} \subset \mathbb{R}^d$ and an output $\mathcal{Y} = \{1, \dots, K\}$ spaces,
- A set of labeled *i.i.d.* training examples $Z_{\mathcal{L}} = (\mathbf{x}_i, y_i)_{1 \le i \le l} \in (\mathcal{X} \times \mathcal{Y})^l$ distributed with respect to a fixed yet unknown probability distribution $\mathcal{D}$ over $\mathcal{X} \times \mathcal{Y}$,
- A set of unlabeled *i.i.d.* training examples $X_{\mathcal{U}} = (\mathbf{x}'_i)_{l+1 \le i \le l+u} \in \mathcal{X}^u$ that are drawn from the marginal distribution $\mathcal{D}_{\mathcal{X}}$ over $\mathcal{X}$,
- A hypothesis space $\mathcal{H}$,
- A posterior distribution Q over $\mathcal{H}$.

We assume that for each $\mathbf{x} \in X_{\mathcal{U}}$ there is exactly one possible label, and $l << u$, which leads to an inefficient supervised model. The *goal* is to minimize an error on the unlabeled set.

Definitions

The Bayes B_Q and the Gibbs G_Q classifiers:

- $B_Q(\mathbf{x}) := \operatorname{argmax}_{c \in \mathcal{Y}} \left[\mathbb{E}_{h \sim Q} \mathbb{1}_{h(\mathbf{x})=c}\right], \quad \forall \mathbf{x} \in \mathcal{X}.$
- G_Q is a stochastic learning algorithm that chooses randomly a hypothesis $h \in \mathcal{H}$ according to the distribution Q and then predicts $h(\mathbf{x})$ for $\mathbf{x} \in \mathcal{X}$.

Transductive measures of error:

- The error rate: $\mathtt{E}_{\mathcal{U}}(h) := \frac{1}{u} \sum_{\mathbf{x}' \in X_{\mathcal{U}}} \mathbb{1}_{h(\mathbf{x}') \neq y'}$,
- The conditional risk: $R_{\mathcal{U}}(h, i, j) := \frac{1}{u_i} \sum_{\mathbf{x}' \in X_{\mathcal{U}}} \mathbb{1}_{h(\mathbf{x}')=j} \mathbb{1}_{y'=i}$,
- The confusion matrix: $\mathbf{C}_h^{\mathcal{U}} = (c_{ij})_{i,j=\{1,\dots,K\}^2}$ with

$$c_{ij} := \begin{cases} 0 & i = j \\ R_{\mathcal{U}}(h, i, j) & i \neq j \end{cases}.$$

where u_i is the size of the class i.
In addition, we consider:

- $m_Q(\mathbf{x}, y) := \mathbb{E}_{h \sim Q} \mathbb{1}_{h(\mathbf{x})=y}$.
- $R_{\mathcal{U} \wedge \boldsymbol{\theta}}(B_Q, i, j) := \frac{1}{u_i} \sum_{\mathbf{x}' \in X_{\mathcal{U}}} \mathbb{1}_{B_Q(\mathbf{x}')=j} \mathbb{1}_{y'=i} \mathbb{1}_{m_Q(\mathbf{x}', j) \ge \theta_j}$
- $\mathtt{E}_{\mathcal{U} \wedge \boldsymbol{\theta}}(B_Q) := \frac{1}{u} \sum_{\mathbf{x}' \in X_{\mathcal{U}}} \mathbb{1}_{B_Q(\mathbf{x}') \neq y'} \mathbb{1}_{m_Q(\mathbf{x}', B_Q(\mathbf{x}')) \ge \theta_{B_Q}(\mathbf{x}')}$.

The error rate and the confusion matrix are connected in the following way:

$$\mathtt{E}_{\mathcal{U}}(h) = \left\|(\mathbf{C}_h^{\mathcal{U}})^\intercal \mathbf{p}\right\|_1, \text{ where } \mathbf{p} = \{u_i/u\}_{i=1}^K.$$

Theorem

Suppose an upper bound $R_u^\delta(G_Q, i, j)$ that holds with prob. $1 - \delta$ is given. Then for any Q and $\forall \delta \in (0, 1]$, $\forall \boldsymbol{\theta} \in [0, 1]^K$ with prob. at least $1 - \delta$ we have:

$$R_{\mathcal{U}}(B_Q, i, j) \le \inf_{\gamma \in [0,1]} \left\{ I_{i,j}^{(\le,<)}(0, \gamma) + \frac{1}{\gamma} \left\lfloor (K_{i,j}^\delta - M_{i,j}^<(\gamma)) \right\rfloor_+ \right\}.$$

$$R_{\mathcal{U} \wedge \boldsymbol{\theta}}(B_Q, i, j) \le \inf_{\gamma \in [\theta_j, 1]} \left\{ I_{i,j}^{(\le,<)}(\theta_j, \gamma) + \frac{1}{\gamma} \left\lfloor (K_{i,j}^\delta - M_{i,j}^<(\gamma) + M_{i,j}^<(\theta_j)) \right\rfloor_+ \right\},$$

where

- $K_{i,j}^\delta = R_u^\delta(G_Q, i, j) - \varepsilon_{i,j}$,
- $\varepsilon_{i,j} = \frac{1}{u_i} \sum_{\mathbf{x}' \in X_{\mathcal{U}}} \mathbb{1}_{B_Q(\mathbf{x}') \neq j} \mathbb{1}_{y'=i} m_Q(\mathbf{x}', j)$,
- $I_{i,j}^{(\le,<)}(\theta_j, \gamma) = \frac{1}{u_i} \sum_{\mathbf{x}' \in X_{\mathcal{U}}} \mathbb{1}_{y'=i} \mathbb{1}_{\theta_j \le m_Q(\mathbf{x}', j) < \gamma}$,
- $M_{i,j}^<(t) = \frac{1}{u_i} \sum_{\mathbf{x}' \in X_{\mathcal{U}}} \mathbb{1}_{y'=i} \mathbb{1}_{m_Q(\mathbf{x}', j) < t} m_Q(\mathbf{x}', j)$.

Corollary

Let $U_{i,j}^\delta(\boldsymbol{\theta})$ be an upper bound for $R_{\mathcal{U} \wedge \boldsymbol{\theta}}(B_Q, i, j)$. Introduce the confusion matrix $\mathbf{U}_{\boldsymbol{\theta}}^\delta$ which (i, j)-entry is 0, if $i = j$, and $U_{i,j}^\delta(\boldsymbol{\theta})$ otherwise. Then, we have:

$$\mathtt{E}_{\mathcal{U} \wedge \boldsymbol{\theta}}(B_Q) \le \left\| \left(\mathbf{U}_{\boldsymbol{\theta}}^\delta\right)^\intercal \mathbf{p} \right\|_1,$$

$$\mathtt{E}_{\mathcal{U}}(B_Q) \le \left\| \left(\mathbf{U}_{\mathbf{0}_K}^\delta\right)^\intercal \mathbf{p} \right\|_1,$$

where $\mathbf{p} = \{u_i/u\}_{i=1}^K$ and $\mathbf{0}_K = (0)_{n=1}^K$.

Algorithm 1: MSLA

Input: Train and unlabelled sets $Z_{\mathcal{L}}$, $X_{\mathcal{U}}$.
A classifier H is trained on $Z_{\mathcal{L}}$.
repeat

1. Compute $\boldsymbol{\theta}^*$ that minimizes the conditional Bayes error rate:

$$\boldsymbol{\theta}^* = \operatorname*{argmin}_{\boldsymbol{\theta} \in (0,1]^K} \mathtt{E}_{\mathcal{U}|\boldsymbol{\theta}}(B_Q).$$

2. From $X_{\mathcal{U}}$ to $Z_{\mathcal{L}}$ move observations $(\mathbf{x}', y')$ such that:

$$[m_Q(\mathbf{x}', y') \ge \theta_{y'}] \wedge [y' = \operatorname*{argmax}_{c \in \mathcal{Y}} m_Q(\mathbf{x}', c)]$$

3. Learn a classifier H on the augmented train set with a new loss:

$$\mathcal{L}(H, Z_{\mathcal{L}}, Z_{\mathcal{U}}) = \frac{l + |Z_{\mathcal{U}}|}{l} \mathcal{L}(H, Z_{\mathcal{L}}) + \frac{l + |Z_{\mathcal{U}}|}{|Z_{\mathcal{U}}|} \mathcal{L}(H, Z_{\mathcal{U}}).$$

until $X_{\mathcal{U}}$ is empty or no add in the train set.
Output: The final classifier H.

Multi-class Self-Learning Algorithm (MSLA)

The principle of MSLA is first to learn a supervised Bayes classifier over the train examples and then iteratively pseudo-labels unlabeled ones for which the margin for the predicted class is no less than a threshold. Then, a new classifier is learned using the train set augmented by pseudo-labeled examples. The process is repeated until there's nothing to add to the train set. At each step, a threshold is found by minimizing the conditional Bayes error rate:

$$\mathtt{E}_{\mathcal{U}|\boldsymbol{\theta}}(B_Q) := \frac{\mathtt{E}_{\mathcal{U} \wedge \boldsymbol{\theta}}(B_Q)}{\frac{1}{u} \sum_{\mathbf{x}' \in X_{\mathcal{U}}} \mathbb{1}_{m_Q(\mathbf{x}', B_Q(\mathbf{x}')) \ge \theta_{B_Q(\mathbf{x}')}}}.$$

Numerical Experiments

We consider 5 datasets, for each of them 20 trials with random train/test split are performed.

Dataset	# of labelled examples, l	# of unlabelled examples, u	Dimension, d	# of classes, K
DNA	31	3155	180	3
MNIST	210	41790	901	10
Pendigits	109	10883	16	10
SensIT	49	22831	100	3
Vowel	99	891	10	11

Table: Description of our experimental setup.

We compare MSLA with the supervised Random Forest (RF) and the multi-class self-learning algorithm with a fixed threshold (FSLA). Both MSLA and FSLA use the Random Forest as the majority vote classifier.

Dataset	Score	RF	MSLA	FSLA $_{\theta=0.7}$	FSLA $_{\theta=0.9}$
DNA	ACC	$.6986 \pm .0767$	$\mathbf{.7076} \pm .0817$	$.5168^{\downarrow} \pm .082$	$.6921 \pm .0752$
	F1	$.6558 \pm .1144$	$\mathbf{.6665} \pm .1174$	$.3747^{\downarrow} \pm .0852$	$.6467 \pm .1141$
MNIST	ACC	$.9039^{\downarrow} \pm .0120$	$\mathbf{.9448} \pm .0061$	$.8654^{\downarrow} \pm .0658$	$.7039^{\downarrow} \pm .0563$
	F1	$.9031^{\downarrow} \pm .0125$	$\mathbf{.9448} \pm .0063$	$.8450^{\downarrow} \pm .0882$	$.6852^{\downarrow} \pm .0647$
Pendigits	ACC	$.861^{\downarrow} \pm .0201$	$\mathbf{.886} \pm .0162$	$.835^{\downarrow} \pm .0384$	$.7998^{\downarrow} \pm .0287$
	F1	$.8586^{\downarrow} \pm .0229$	$\mathbf{.8845} \pm .0171$	$.8257^{\downarrow} \pm .0488$	$.7906^{\downarrow} \pm .0358$
SensIT	ACC	$.67 \pm .0291$	$\mathbf{.6745} \pm .0288$	$.6192^{\downarrow} \pm .0366$	$.53^{\downarrow} \pm .0391$
	F1	$.654 \pm .0448$	$\mathbf{.6599} \pm .0421$	$.5784^{\downarrow} \pm .0683$	$.4302^{\downarrow} \pm .0887$
Vowel	ACC	$\mathbf{.5851} \pm .0273$	$.5846 \pm .0268$	$.5265^{\downarrow} \pm .0374$	$.5839 \pm .0292$
	F1	$.5733 \pm .0293$	$\mathbf{.5754} \pm .0278$	$.5053^{\downarrow} \pm .0407$	$.5713 \pm .0311$

Table: Classification performance on different datasets described in Table 1. Two score functions are computed, namely, accuracy and F1. The sign $^{\downarrow}$ shows if the performance is significantly worse than the best result on the level 0.01.

Results

- Overall, the MSLA performs better than the others. For the MNIST and the Pendigits datasets the improvement is reported as significant.
- One can notice that regardless the possible benefit MSLA could provide, there is always an unrecoverable error that the basis classifier produces on the initial step of the MSLA.
- In our experiments we have not found a case when the FSLA has any benefit, since it performs worse than the supervised approach.

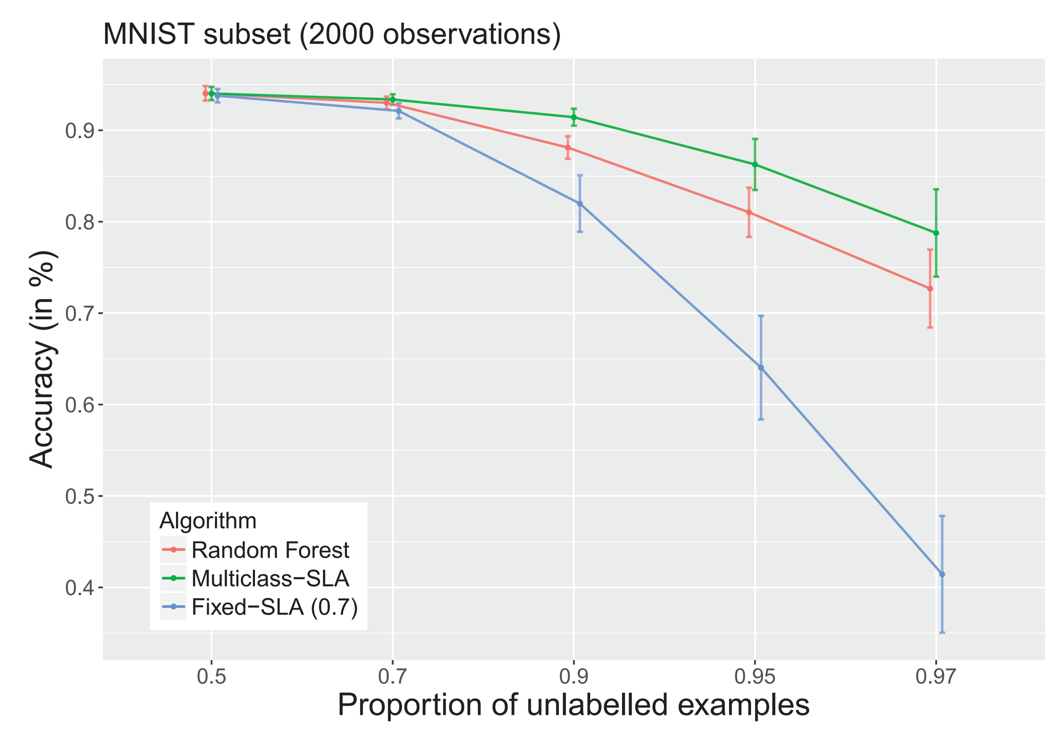

Figure: Classification accuracy w.r.t. the proportion of unlabeled examples for the MNIST dataset.

References

[1] Massih-Reza Amini, François Laviolette, and Nicolas Usunier.
A transductive bound for the voted classifier with an application to semi-supervised learning.
In *Advances in NIPS 21, Proceedings of the 22nd Annual Conf. on NIPS, Vancouver, Canada, Dec. 8-11, 2008*, pages 65–72, 2008.

[2] Vasilii Feofanov, Émilie Devijver, and Massih-Reza Amini.
Une borne transductive multiclasses pour le classifieur de vote majoritaire
(work in progress).