Transductive Bounds for the Multi-class Majority Vote Classifier

Vasilii Feofanov, Emilie Devijver, Massih-Reza Amini

▶ **To cite this version:**

Vasilii Feofanov, Emilie Devijver, Massih-Reza Amini. Transductive Bounds for the Multi-class Majority Vote Classifier. Proceedings of the AAAI Conference on Artificial Intelligence, 2019, 33, pp.3566-3573. 10.1609/aaai.v33i01.33013566 . hal-01980449v2

HAL Id: hal-01980449

https://hal.archives-ouvertes.fr/hal-01980449v2

Submitted on 4 Jan 2021

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Transductive Bounds for the Multi-Class Majority Vote Classifier

Vasilii Feofanov, Emilie Devijver, Massih-Reza Amini
University Grenoble Alpes, Grenoble INP
LIG, CNRS, Grenoble 38000, France
firstname.lastname@univ-grenoble-alpes.fr

Abstract

In this paper, we propose a transductive bound over the risk of the majority vote classifier learned with partially labeled data for the multi-class classification. The bound is obtained by considering the class confusion matrix as an error indicator and it involves the margin distribution of the classifier over each class and a bound over the risk of the associated Gibbs classifier. When this latter bound is tight and, the errors of the majority vote classifier per class are concentrated on a low margin zone; we prove that the bound over the Bayes classifier' risk is tight. As an application, we extend the self-learning algorithm to the multi-class case. The algorithm iteratively assigns pseudo-labels to a subset of unlabeled training examples that have their associated class margin above a threshold obtained from the proposed transductive bound. Empirical results on different data sets show the effectiveness of our approach compared to the same algorithm where the threshold is fixed manually, to the extension of TSVM to multi-class classification and to a graph-based semi-supervised algorithm.

1 Introduction

In many real-life applications, the labeling of training examples for learning is costly and sometimes even not realistic. For example, in medical diagnosis or biological data analysis, labeling data may require very expensive tests so that only small labeled data sets are generally available. In many other cases, like web oriented applications, huge amount of observations arrive sequentially and there is not enough time to label data for different information needs; while unlabeled data are abundant.

Learning with labeled and unlabeled data, or semi-supervised learning, has been a subject of growing interest in the machine learning community over the last twenty years (Chapelle, Schölkopf, and Zien 2010). In this case, labeled training examples are generally assumed to be very few, leading to an inefficient supervised model, while unlabeled training examples contain valuable information about the prediction problem and it is generally expected that their exploitation leads to an increase of prediction performance.

Considering an input space $\mathcal{X} \subset \mathbb{R}^d$ and a discrete output space $\mathcal{Y}$, we assume available a set of labeled training examples $\mathrm{Z}_{\mathcal{L}} = \{(\mathbf{x}_i, y_i)\}_{i=1}^{l} \in (\mathcal{X} \times \mathcal{Y})^l$, identically and independently distributed (i.i.d.) with respect to a fixed yet unknown probability distribution $\mathcal{D}$ over $\mathcal{X} \times \mathcal{Y}$, and a set of unlabeled training examples $\mathrm{X}_{\mathcal{U}} = \{\mathbf{x}'_i\}_{i=l+1}^{l+u} \in \mathcal{X}^u$ supposed to be drawn i.i.d. from the marginal distribution $P_X(\mathbf{x})$, over the domain $\mathcal{X}$. If $X_{\mathcal{U}}$ is empty, then the problem reduces to supervised learning. The other extreme case is the situation where $\mathrm{Z}_{\mathcal{L}}$ is empty and which corresponds to unsupervised learning.

Most studies in semi-supervised learning have focused on the binary classification problem, whereas just few ones are devoted to the multi-class framework, i.e. $|\mathcal{Y}| > 2$ with some recent studies considering the learnability of multi-class semi-supervised learning algorithms under some specific assumptions. For example, Maximov, Amini, and Harchaoui (2018) proved the consistency of the Empirical Risk Minimization principle in some cases by bounding the true risk of the trained classifier. However such bounds are not usable in practice as they are generally too loose.

In this paper, we propose a transductive bound for the multi-class majority vote classifier, which to the best of our knowledge, is a first attempt in this direction. The bound is based on the risk of the associated Gibbs classifier and by considering the class confusion matrix as an error indicator as proposed in Morvant, Koço, and Ralaivola (2012). This bound is obtained by analytically solving a linear program and it comes out that in the case where the bound over the risk of the Gibbs classifier is tight and when the Bayes classifier makes most of its errors on low margin examples, the obtained bound is tight. From this result, we then propose to automatically find a threshold for which the risk of the majority vote classifier is the lowest. This finding allows to consider the output of the Bayes classifier, or its margin, as an indicator of confidence and to extend self-learning algorithms to the multi-class case. The proposed strategy iteratively learns a Bayes classifier by assigning at each iteration pseudo-labels to unlabeled examples having their margin above a certain threshold obtained from the proposed transductive bound.

The paper is organized as follows. In Section 2 we introduce the problem statement and the proposed framework. In Section 3 we present a bound over the transductive risk of the multi-class majority vote classifier. In Section 4 we present empirical evidence showing that the extended self-

Copyright © 2019, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

learning algorithm learned using the proposed bound is effective compared to the same algorithm where the threshold is fixed manually, to the extension of TSVM to multi-class classification and to a graph-based semi-supervised algorithm on difference data sets. Finally, in Section 5 we discuss the outcomes of this study and give some pointers to further research.

2 Framework and Definitions

In this study, we consider learning algorithms that work in a fixed hypothesis space $\mathcal{H} = \{h : \mathcal{X} \to \mathcal{Y}\}$ of multi-class classifiers (defined without reference to the training data). After observing the training set $S = \mathrm{Z}_{\mathcal{L}} \cup \mathrm{X}_{\mathcal{U}}$, the task of the learner is then to choose a posterior distribution Q over $\mathcal{H}$ such that the Q-weighted majority vote classifier (also called Bayes classifier)

$$B_Q(\mathbf{x}) := \operatorname*{argmax}_{c\in\mathcal{Y}} \left[\mathbb{E}_{h\sim Q}\mathbb{1}_{h(\mathbf{x})=c}\right], \quad \forall \mathbf{x} \in \mathcal{X}, \qquad (1)$$

will have the smallest possible risk on examples of $\mathrm{X}_{\mathcal{U}}$. Together with that, we consider the associated Gibbs classifier G_Q that for any $\mathbf{x} \in \mathcal{X}$ chooses randomly a classifier $h \in \mathcal{H}$ according to Q. We accordingly define the *transductive error rate*of B_Q and G_Q over an unlabeled set by:

$$\mathtt{E}_{\mathcal{U}}(B_Q) := \frac{1}{u}\sum_{\mathbf{x}'\in\mathrm{X}_{\mathcal{U}}} \mathbb{1}_{B_Q(\mathbf{x}')\neq y'}, \qquad (2)$$

$$\mathtt{E}_{\mathcal{U}}(G_Q) := \frac{1}{u}\sum_{\mathbf{x}'\in\mathrm{X}_{\mathcal{U}}} \mathbb{E}_{h\sim Q}\mathbb{1}_{h(\mathbf{x}')\neq y'}. \qquad (3)$$

For an observation $\mathbf{x}$, we further define its *unsigned margin* $\mathbf{m}_{\mathbf{x}} = (m_Q(\mathbf{x}, c))_{c=1}^K$ which measures the confidence in each class of the classifier as

$$m_Q(\mathbf{x}, y) := \mathbb{E}_{h\sim Q}\mathbb{1}_{h(\mathbf{x})=y}. \qquad (4)$$

The proposed bound follows a bound on the *joint Bayes error rate* which given a vector $\boldsymbol{\theta} = (\theta_n)_{n=1}^K \in [0, 1]^K$, is defined as,

$$\mathtt{E}_{\mathcal{U}\wedge\boldsymbol{\theta}}(B_Q) := \frac{1}{u}\sum_{\mathbf{x}'\in\mathrm{X}_{\mathcal{U}}} \mathbb{1}_{B_Q(\mathbf{x}')\neq y'}\mathbb{1}_{m_Q(\mathbf{x}', B_Q(\mathbf{x}'))\geq\theta_k},$$

where y' is the true unknown class label of $\mathbf{x}'$. One of the practical issues that arises from this result is the possibility to define a set of thresholds $\boldsymbol{\theta}$ for which the bound is optimal and that we use in a self-learning algorithm by iteratively assigning pseudo-labels to unlabeled examples having the highest class margin above the corresponding threshold.

However, as we work with multi-class data, the error rate does not describe the dispersion of errors regarding each class over all the others. We rather use the confusion matrix, which provides a richer information. For a classifier h, the *transductive confusion matrix* $\mathbf{C}_h^{\mathcal{U}} = (c_{ij})_{i,j=\{1,\dots,K\}^2}$ is defined as follows:

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

where for a classifier h, for each class pair $(i, j) \in \{1, \dots, K\}^2$ s.t. $i \neq j$, the *transductive conditional risk* $R_{\mathcal{U}}$ is defined by:

$$R_{\mathcal{U}}(h, i, j) := \frac{1}{u_i}\sum_{\mathbf{x}'\in\mathrm{X}_{\mathcal{U}}} \mathbb{1}_{h(\mathbf{x}')=j}\mathbb{1}_{y'=i},$$

with $u_i = \sum_{\mathbf{x}'\in\mathrm{X}_{\mathcal{U}}} \mathbb{1}_{y'=i}$ is the size of class $i \in \mathcal{Y}$.

Similarly, the *transductive conditional Gibbs risk* is defined as $R(G_Q, i, j) := \mathbb{E}_{h\sim Q}R(h, i, j)$.

The *transductive joint Bayes confusion matrix* $\mathbf{C}_{B_Q}^{\mathcal{U}\wedge\boldsymbol{\theta}} = (c_{ij})_{i,j=\{1,\dots,K\}^2}$ given a vector $\boldsymbol{\theta} = (\theta_n)_{n=1}^K$, $\boldsymbol{\theta} \in [0, 1]^K$ is defined as:

$$c_{ij} := \begin{cases} 0 & i = j, \\ R_{\mathcal{U}\wedge\boldsymbol{\theta}}(B_Q, i, j) & i \neq j, \end{cases}$$

where the *transductive joint Bayes conditional risk* $R_{\mathcal{U}\wedge\boldsymbol{\theta}}(B_Q, i, j)$ for the class pair $(i, j) \in \{1, \dots, K\}^2$ s.t. $i \neq j$, is defined as follows:

$$R_{\mathcal{U}\wedge\boldsymbol{\theta}}(B_Q, i, j) := \frac{1}{u_i}\sum_{\mathbf{x}'\in\mathrm{X}_{\mathcal{U}}} \mathbb{1}_{B_Q(\mathbf{x}')=j}\mathbb{1}_{y'=i}\mathbb{1}_{m_Q(\mathbf{x}',j)\geq\theta_j}.$$

Thus, the transductive joint Bayes conditional risk counts an example as wrongly classified, if its true label is i and the majority vote classifier predicts the class j with the margin $m_Q(\mathbf{x}', j) \geq \theta_j$. Generally, the majority vote classifier is supposed to make errors by predicting the label j mostly on examples with a low value of $m_Q(\mathbf{x}', j)$. Then, if θ_j is high enough, the joint conditional risk computes the probability to make a mistake on high margin "confident" observations.

To work with matrices, we use the *spectral norm*, defined by, for a matrix $\mathbf{A}$ of size $n \times m$:

$$\|\mathbf{A}\|_2 := \sup_{\substack{\mathbf{x}\in\mathbb{R}^m \\ \|\mathbf{x}\|_2=1}} \|\mathbf{A}\mathbf{x}\|_2 = \sup_{\mathbf{x}\in\mathbb{R}^m} \frac{\|\mathbf{A}\mathbf{x}\|_2}{\|\mathbf{x}\|_2}.$$

It corresponds to the matrix's largest singular value.

We conclude this section by the following proposition, which links the error rate to the confusion matrix.

Proposition 1. *Let B_Q be the Bayes classifier. Given a vector $\boldsymbol{\theta} \in [0, 1]^K$, for $\mathbf{p} := \{u_i/u\}_{i=1}^K$, we have:*

$$\mathtt{E}_{\mathcal{U}\wedge\boldsymbol{\theta}}(B_Q) = \left\| \left(\mathbf{C}_{B_Q}^{\mathcal{U}\wedge\boldsymbol{\theta}}\right)^{\intercal} \mathbf{p} \right\|_1.$$

3 Transductive Bounds on the Risk of the Multi-class Majority Vote Classifier

In this section we propose a transductive bound for the majority vote classifier in multi-class setting. The bound is based on the margin distribution as well as a bound of the transductive conditional Gibbs risk, which we suppose given. First, we give a theorem that provides a bound for the transductive joint Bayes conditional risk, which leads to a bound for the transductive conditional risk of the majority vote classifier. Then, a corollary is derived, proposing upper bounds for the Bayes confusion matrix and the Bayes error rate. Finally, we propose a setting under which the bound on the conditional risk of the Bayes classifier becomes tight.

Main Result

Theorem 1. *Let B_Q be the Q-weighted majority vote classifier. Suppose an upper bound of the transductive conditional Gibbs risk $R_u^\delta(G_Q, i, j)$ that holds with probability $1-\delta$ is given. Then for any Q and $\forall \delta \in (0,1], \forall \boldsymbol{\theta} \in [0,1]^K$, $\forall (i,j) \in \mathcal{Y}^2$, with probability at least $1-\delta$ we have:*

$$R_{\mathcal{U}}(B_Q, i, j) \leq \inf_{\gamma \in [0,1]} \Big\{ I_{i,j}^{(\leq,<)}(0,\gamma) + \frac{1}{\gamma} \big\lfloor (K_{i,j}^\delta - M_{i,j}^<(\gamma)) \big\rfloor_+ \Big\}, \quad (5)$$

$$R_{\mathcal{U}\wedge\boldsymbol{\theta}}(B_Q, i, j) \leq \inf_{\gamma \in [\theta_j, 1]} \Big\{ I_{i,j}^{(\leq,<)}(\theta_j,\gamma) + \frac{1}{\gamma} \big\lfloor (K_{i,j}^\delta - M_{i,j}^<(\gamma) + M_{i,j}^<(\theta_j)) \big\rfloor_+ \Big\}, \quad (6)$$

where

$$K_{i,j}^\delta = R_u^\delta(G_Q, i, j) - \varepsilon_{i,j},$$
$$\varepsilon_{i,j} = \frac{1}{u_i} \sum_{\mathbf{x}' \in \mathrm{X}_{\mathcal{U}}} \mathbb{1}_{B_Q(\mathbf{x}') \neq j} \mathbb{1}_{y'=i} m_Q(\mathbf{x}', j),$$
$$I_{i,j}^{(\lhd_1,\lhd_2)}(t,s) = \frac{1}{u_i} \sum_{\mathbf{x}' \in \mathrm{X}_{\mathcal{U}}} \mathbb{1}_{y'=i} \mathbb{1}_{t \lhd_1 m_Q(\mathbf{x}', j) \lhd_2 s},$$
$$(\lhd_1, \lhd_2) \in \{<, \leq\}^2,$$
$$M_{i,j}^<(t) = \frac{1}{u_i} \sum_{\mathbf{x}' \in \mathrm{X}_{\mathcal{U}}} \mathbb{1}_{y'=i} \mathbb{1}_{m_Q(\mathbf{x}', j) < t} \, m_Q(\mathbf{x}', j),$$
$$\lfloor x \rfloor_+ = x \cdot \mathbb{1}_{x>0},$$

From spectral norm properties, the following corollary is easily deduced:

Corollary 1. *Let $U_{i,j}^\delta(\boldsymbol{\theta})$ be the upper bound for the transductive joint Bayes conditional risk from Theorem 1 that holds $\forall (i,j) \in \mathcal{Y}^2, \forall \delta \in (0,1], \forall \boldsymbol{\theta} \in [0,1]^K$, with probability at least $1-\delta$:*

$$U_{i,j}^\delta(\boldsymbol{\theta}) := \inf_{\gamma \in [\theta_j, 1]} \Big\{ I_{i,j}^{(\leq,<)}(\theta_j,\gamma) + \frac{1}{\gamma} \big\lfloor (K_{i,j}^\delta - M_{i,j}^<(\gamma) + M_{i,j}^<(\theta_j)) \big\rfloor_+ \Big\}. \quad (7)$$

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. We consider the spectral norm. Then, we have:

$$\|\mathbf{C}_{B_Q}^{\mathcal{U}\wedge\boldsymbol{\theta}}\| \leq \|\mathbf{U}_{\boldsymbol{\theta}}^\delta\| \text{ and } \|\mathbf{C}_{B_Q}^{\mathcal{U}}\| \leq \|\mathbf{U}_{\mathbf{0}_K}^\delta\|,$$

where $\mathbf{0}_K$ is the K-size vector of zeros.

Moreover, we have:

$$\mathtt{E}_{\mathcal{U}\wedge\boldsymbol{\theta}}(B_Q) \leq \left\| (\mathbf{U}_{\boldsymbol{\theta}}^\delta)^\intercal \, \mathbf{p} \right\|_1 \text{ and } \mathtt{E}_{\mathcal{U}}(B_Q) \leq \left\| (\mathbf{U}_{\mathbf{0}_K}^\delta)^\intercal \, \mathbf{p} \right\|_1,$$

where $\mathbf{p} = \{u_i/u\}_{i=1}^K$.

In the following proposition, we assume that the classifier makes most of its error on unlabeled examples with low margin. Then, considering that the margin is an indicator of confidence, the bound becomes tight.

Proposition 2. *For all $\mathbf{x}' \in \mathrm{X}_{\mathcal{U}}$ there exists $C \in [0,1]$ such that for all $(i,j) \in \mathcal{Y}^2$, for all $\gamma > 0$:*

$$\frac{1}{u_i} \sum_{\mathbf{x}' \in \mathrm{X}_{\mathcal{U}}} \mathbb{1}_{B_Q(\mathbf{x}')=j} \mathbb{1}_{y'=i} \mathbb{1}_{m_Q(\mathbf{x}', j)=\gamma} \neq 0 \Rightarrow$$
$$\frac{1}{u_i} \sum_{\mathbf{x}' \in \mathrm{X}_{\mathcal{U}}} \mathbb{1}_{B_Q(\mathbf{x}')=j} \mathbb{1}_{y'=i} \mathbb{1}_{m_Q(\mathbf{x}', j)<\gamma} \geq C \cdot \frac{1}{u_i} \sum_{\mathbf{x}' \in \mathrm{X}_{\mathcal{U}}} \mathbb{1}_{y'=i} \mathbb{1}_{m_Q(\mathbf{x}', j)<\gamma}. \quad (8)$$

Then, with probability at least $1-\delta$ the following inequality holds:

$$F_{i,j}^\delta - R_{\mathcal{U}}(B_Q, i, j) \leq \frac{1-C}{C} R_{\mathcal{U}}(B_Q, i, j) + \frac{R_u^\delta(G_Q, i, j) - R_{\mathcal{U}}(G_Q, i, j)}{\gamma^*}, \quad (9)$$

where

- $\gamma^* := \gamma^{(p)}$, *where* $p := \sup \Big\{ w \in \{1, \ldots, N_j\} | b_{i,j}^{(w)} \neq 0 \Big\}$.
- $F_{i,j}^\delta := \inf_{\gamma \in [0,1]} \Big\{ I_{i,j}^{(\leq,<)}(0,\gamma) + \frac{1}{\gamma} \big\lfloor (K_{i,j}^\delta - M_{i,j}^<(\gamma)) \big\rfloor_+ \Big\}$.

In the next section, proofs are provided.

Proofs

Proof of Theorem 1. This proof relies on two lemmas. The first one connects the conditional Gibbs risk and the conditional joint Bayes risk. The second one provides an analytic solution of a linear program.

Lemma 1. *Let $\Gamma_c := \{\gamma_c | \, \exists \, \mathbf{x}' \in \mathrm{X}_{\mathcal{U}} : \gamma_c = m_Q(\mathbf{x}', c)\}$, where $c \in \mathcal{Y}$, and $N_c := |\Gamma_c|$. Let enumerate its elements such that they form an ascending order:*

$$\gamma_c^{(1)} \leq \gamma_c^{(2)} \leq \cdots \leq \gamma_c^{(N_c)}.$$

Denote $b_{i,j}^{(n)} := \frac{1}{u_i} \sum_{\mathbf{x}' \in \mathrm{X}_{\mathcal{U}}} \mathbb{1}_{B_Q(\mathbf{x}')=j} \mathbb{1}_{y'=i} \mathbb{1}_{m_Q(\mathbf{x}', j)=\gamma_j^{(n)}}$. Then, $\forall (i,j) \in \mathcal{Y}^2$:

$$R_{\mathcal{U}}(G_Q, i, j) = \sum_{n=1}^{N_j} b_{i,j}^{(n)} \gamma_j^{(n)} + \varepsilon_{i,j}, \quad (10)$$

$$R_{\mathcal{U}\wedge\boldsymbol{\theta}}(B_Q, i, j) = \sum_{n=k+1}^{N_j} b_{i,j}^{(n)}, \quad (11)$$

where $\varepsilon_{i,j} = \frac{1}{u_i} \sum_{\mathbf{x}' \in \mathrm{X}_{\mathcal{U}}} \mathbb{1}_{B_Q(\mathbf{x}') \neq j} \mathbb{1}_{y'=i} m_Q(\mathbf{x}', j)$ and

$$k = \begin{cases} 0 & \text{if } \{n | \gamma_j^{(n)} < \theta_j\} = \emptyset \\ \max\{n | \gamma_j^{(n)} < \theta_j\} & \text{otherwise.} \end{cases}$$

Proof. Formula (10) is derived through conditioning by the value of the majority vote classifier.

Formula (11) is get by considering k, the index of the smallest γ larger than the threshold θ:

$$R_{\mathcal{U}\wedge\boldsymbol{\theta}}(B_Q, i, j) = \frac{1}{u_i} \sum_{n=k+1}^{N_j} \sum_{\mathbf{x}' \in X_{\mathcal{U}}} \mathbb{1}_{B_Q(\mathbf{x}')=j} \mathbb{1}_{y'=i} \mathbb{1}_{m_Q(\mathbf{x}', j)=\gamma_j^{(n)}}.$$

We conclude by definition of $b_{i,j}^{(n)}$.

□

Lemma 2 (Lemma 4 in Amini, Laviolette, and Usunier (2008))**.** *Let $(g_i)_{i\in\{1,\dots,N\}}$ be such that $0 < g_1 < g_2 < \cdots < g_{N-1} < g_N \leq 1$. Consider also $p_i \geq 0$, $i = 1,\dots,N$, $B \geq 0$, $k \in \{1,\dots,N\}$. Then, the optimal solution of the linear program:*

$$\begin{cases} \max_{q:=(q_1,\dots,q_N)} F(q) := \max_{q_1,\dots,q_N} \sum_{i=k+1}^{N} q_i \\ 0 \leq q_i \leq p_i \quad \forall i \in \{1,\dots,N\} \\ \sum_{i=1}^{N} q_i g_i \leq B \end{cases}$$

will be q^* defined as $\forall i \leq k : \; q_i^* = 0$, $\forall i > k : \; q_i^* = \min\left(p_i, \left\lfloor \frac{B-\sum_{j<i} q_j^* g_j}{g_i} \right\rfloor_+\right)$.

Now we combine those two lemmas to prove Theorem 1.

First, notice that Eq. (5) is easily derived from Eq. (6) using that $M_{i,j}^{<}(0) = 0$.

To prove Eq. (6), we consider two cases.

First, $\forall (i,j)$, $\forall \boldsymbol{\theta} \in [0,1]^K$, when the mistake is maximized, using Lemma 1, we get:

$$R_{\mathcal{U}\wedge\boldsymbol{\theta}}(B_Q,i,j) = \sum_{n=k}^{N_j} b_{i,j}^{(n)} \leq \max_{b_{i,j}^{(1)},\dots,b_{i,j}^{(N_j)}} \sum_{n=k}^{N_j} b_{i,j}^{(n)}, \quad (12)$$

with k is equal to 0 when $\{n|\gamma_j^{(n)} < \theta_j\} = \emptyset$, and $\max\{n|\gamma_j^{(n)} < \theta_j\}$ otherwise.

Consider the upper bound $R_u^\delta(G_Q,i,j)$ of the Gibbs conditional risk $R_{\mathcal{U}}(G_Q,i,j)$ that holds with probability $1-\delta$. Denote $K_{i,j}^\delta = R_u^\delta(G_Q,i,j) - \varepsilon_{i,j}$ and $B_{i,j}^{(n)} = \frac{1}{u_i}\sum_{\mathbf{x}'\in \mathrm{X}_{\mathcal{U}}} \mathbb{1}_{y'=i}\mathbb{1}_{m_Q(\mathbf{x}',j)=\gamma_j^{(n)}}$. We are interested in the following linear program task:

$$\max_{b_{i,j}^{(1)},\dots,b_{i,j}^{(N_j)}} \sum_{n=k}^{N_j} b_{i,j}^{(n)}$$

$$\text{s.t. } \forall n,\ 0 \leq b_{i,j}^{(n)} \leq B_{i,j}^{(n)} \text{ and } \sum_{n=1}^{N_j} b_{i,j}^{(n)}\gamma_j^{(n)} \leq K_{i,j}^\delta. \quad (13)$$

As $\sum_{k<w<n} \gamma_j^{(w)} B_{i,j}^{(w)} = M_{i,j}^{<}(\gamma_j^{(n)}) - M_{i,j}^{<}(\theta_j)$ with $k = \max\{w|\gamma_j^{(w)} < t\}$, we get the following solution of the linear program (13) by using Lemma 2: with $p = \max\{n|K_{i,j}^\delta - M_{i,j}^{<}(\gamma_j^{(n)}) + M_{i,j}^{<}(\theta_j) > 0\}$,

$$b_{i,j}^{(n)} = \begin{cases} 0 & n \leq k, \\ B_{i,j}^{(n)} & n \in [k+1, p), \\ \frac{1}{\gamma_j^{(p)}}(K_{i,j}^\delta - M_{i,j}^{<}(\gamma_j^{(p)}) + M_{i,j}^{<}(\theta_j)) & n = p, \\ 0 & n > p. \end{cases}$$

This formulae is used to rewrite Eq. (12), as $\sum_{n=k+1}^{p-1} B_{i,j}^{(n)} = I_{i,j}^{(\leq,<)}(\theta_j,\gamma)$:

$$R_{\mathcal{U}\wedge\boldsymbol{\theta}}(B_Q,i,j) \leq I_{i,j}^{(\leq,<)}(\theta_j,\gamma_j^{(p)}) + \frac{1}{\gamma_j^{(p)}}(K_{i,j}^\delta - M_{i,j}^{<}(\gamma_j^{(p)}) + M_{i,j}^{<}(\theta_j)).$$

Consider the function: $\gamma \mapsto T_{i,j}(\gamma) := I_{i,j}^{(\leq,<)}(\theta_j,\gamma) + \frac{1}{\gamma}\left\lfloor (K_{i,j}^\delta - M_{i,j}^{<}(\gamma) + M_{i,j}^{<}(\theta_j))\right\rfloor_+$. To prove the theorem, it remains to check that $\forall \gamma \in [\theta_j, 1]$, $T_{i,j}(\gamma_j^{(p)}) \leq T_{i,j}(\gamma)$. For this, let's consider $\gamma_j^{(w)}$, $w \in \{1,\dots,N_j\}$.

If $w > p$, then

$$T_{i,j}(\gamma_j^{(p)}) \leq I_{i,j}^{(\leq,<)}(\theta_j,\gamma_j^{(p)}) \leq T_{i,j}(\gamma_j^{(w)}).$$

If $w < p$, then

$$\begin{aligned} &T(\gamma_j^{(p)}) - T(\gamma_j^{(w)}) \\ &= \sum_{n=w}^{p} b_{i,j}^{(n)} - \frac{1}{\gamma_j^{(w)}}(K_{i,j}^\delta - M_{i,j}^{<}(\gamma_j^{(w)}) + M_{i,j}^{<}(\theta_j)) \\ &= \sum_{n=w}^{p} b_{i,j}^{(n)} - \frac{1}{\gamma_j^{(w)}}\Big(\sum_{n=k+1}^{p} b_{i,j}^{(n)}\gamma_j^{(n)} - \sum_{n=k+1}^{w-1} \gamma_j^{(n)} b_{i,j}^{(n)}\Big) \\ &= \frac{1}{\gamma_j^{(w)}}\Big(\sum_{n=w}^{p} b_{i,j}^{(n)}\gamma_j^{(w)} - \sum_{n=w}^{p} b_{i,j}^{(n)}\gamma_j^{(n)}\Big) \leq 0. \end{aligned}$$

Summing up, we derive for $R_{\mathcal{U}\wedge\boldsymbol{\theta}}(B_Q,i,j)$ the upper bound $T_{i,j}(\gamma_j^{(p)})$, which, in addition, is the infimum of $T_{i,j}$ on $\gamma \in [\theta_j, 1]$.

□

Proof of Proposition 2. First, let's show that, $\forall (i,j) \in \mathcal{Y}^2$,

$$R_{\mathcal{U}}(B_Q,i,j) \geq \frac{1}{u_i}\sum_{\mathbf{x}'\in \mathrm{X}_{\mathcal{U}}} \mathbb{1}_{B_Q(\mathbf{x}')=j}\mathbb{1}_{y'=i}\mathbb{1}_{m_Q(\mathbf{x}',j)<\gamma^*} + \frac{1}{\gamma^*}\lfloor K_{i,j} - M_{i,j}^{<}(\gamma^*)\rfloor_+, \quad (14)$$

where $K_{i,j} = R_{\mathcal{U}}(G_Q,i,j) - \varepsilon_{i,j}$ and $\varepsilon_{i,j} = \frac{1}{u_i}\sum_{\mathbf{x}'\in \mathrm{X}_{\mathcal{U}}} \mathbb{1}_{B_Q(\mathbf{x}')\neq j}\mathbb{1}_{y'=i} m_Q(\mathbf{x}',j)$.

Denote $\gamma^* = \gamma_j^{(p)}$. We apply Lemma 1 and get that $R_{\mathcal{U}}(G_Q,i,j) = \sum_{n=1}^{p} b_{i,j}^{(n)}\gamma_j^{(n)} + \varepsilon_{i,j}$, where $b_{i,j}^{(n)} := \frac{1}{u_i}\sum_{\mathbf{x}'\in \mathrm{X}_{\mathcal{U}}} \mathbb{1}_{B_Q(\mathbf{x}')=j}\mathbb{1}_{y'=i}\mathbb{1}_{m_Q(\mathbf{x}',j)=\gamma_j^{(n)}}$. Then, we can write $b_{i,j}^{(p)} = \left(K_{i,j} - \sum_{n=1}^{p-1} b_{i,j}^{(n)}\gamma_j^{(n)}\right)/\gamma_j^{(p)}$. Since $-\sum_{n=1}^{p-1} b_{i,j}^{(n)}\gamma_j^{(n)} \geq -M_{i,j}^{<}(\gamma_j^{(p)})$, we deduce a lower bound for $b_{i,j}^{(p)}$:

$$b_{i,j}^{(p)} \geq \frac{1}{\gamma^*}\lfloor K_{i,j} - M_{i,j}^{<}(\gamma^*)\rfloor_+. \quad (15)$$

Also, taking into account Lemma 1, one can notice that $R_{\mathcal{U}}(B_Q,i,j) = \sum_{n=1}^{p} b_{i,j}^{(n)} = b_{i,j}^{(p)} + \frac{1}{u_i}\sum_{\mathbf{x}'\in \mathrm{X}_{\mathcal{U}}} \mathbb{1}_{B_Q(\mathbf{x}')=j}\mathbb{1}_{y'=i}\mathbb{1}_{m_Q(\mathbf{x}',j)<\gamma^*}$. Combining this fact and Eq. (15) we deduce Eq. (14).

Using the initial assumptions and following the definition of $I_{i,j}^{(\lhd_1,\lhd_2)}(t,s)$ we deduce from Eq. (14):

$$R_{\mathcal{U}}(B_Q,i,j) \geq C \cdot I_{i,j}^{(\leq,<)}(0,\gamma^*) + \frac{1}{\gamma^*}\lfloor K_{i,j} - M_{i,j}^{<}(\gamma^*)\rfloor_+. \quad (16)$$

Notice that $F^{\delta}_{i,j} \leq I^{(\leq,<)}_{i,j}(0,\gamma^*) + \frac{1}{\gamma^*}\left\lfloor (K^{\delta}_{i,j} - M^{<}_{i,j}(\gamma^*))\right\rfloor_+$. Subtracting Eq. (16) from this inequality we obtain:

$$F^{\delta}_{i,j} - R_{\mathcal{U}}(B_Q,i,j) \leq (1-C) I^{(\leq,<)}_{i,j}(0,\gamma^*) + \frac{1}{\gamma^*}\left(\left\lfloor (K^{\delta}_{i,j} - M^{<}_{i,j}(\gamma^*))\right\rfloor_+ - \lfloor K_{i,j} - M^{<}_{i,j}(\gamma^*)\rfloor_+\right),$$

which holds with probability $1-\delta$.

Then, as by definition $R^{\delta}_{\mathcal{U}}(G_Q) \geq R_{\mathcal{U}}(G_Q)$ holds with probability $1-\delta$, we obtain:

$$K^{\delta}_{i,j} - K_{i,j} = R^{\delta}_{\mathcal{U}}(G_Q,i,j) - R_{\mathcal{U}}(G_Q,i,j) \geq 0. \quad (17)$$

In addition, since for any non-negative real numbers $a,b,m \in \mathbb{R}^+$ with $b \geq a$, it is true that $\lfloor b-m\rfloor_+ - \lfloor a-m\rfloor_+ \leq b-a$, we deduce that

$$\left\lfloor (K^{\delta}_{i,j} - M^{<}_{i,j}(\gamma^*))\right\rfloor_+ - \left\lfloor K_{i,j} - M^{<}_{i,j}(\gamma^*)\right\rfloor_+ \leq R^{\delta}_{\mathcal{U}}(G_Q,i,j) - R_{\mathcal{U}}(G_Q,i,j). \quad (18)$$

Finally, from Eq. (16) we derive that $I^{(\leq,<)}_{i,j}(0,\gamma^*) \leq \frac{1}{C} R_{\mathcal{U}}(B_Q,i,j)$. Taking into account this fact as well as Eq. (18), we infer Eq. (9).

□

Proposition 2 states that if Condition (8) holds, the difference between the conditional Bayes risk and its upper bound does not exceed an expression that depends on a constant C. If we assume that the Gibbs conditional risk bound is as tight as possible and the majority vote classifier makes most of its mistake for the class j on observations with the low value of $m_Q(\mathbf{x}',j)$, we obtain that Condition (8) accepts a high value C (close to 1), and the bound becomes tight. From theoretical point of view it makes sense to assume that the majority vote classifier mistakes mostly on low margin region, since if the class got a relatively high vote from the hypotheses, we expect that it is predicted correctly.

4 Multi-class Self-Learning Algorithm

Algorithm

Pseudo-labelling is considered in this paper to increase the labeled set and improve performances. We introduce the *conditional Bayes error rate* $\mathtt{E}_{\mathcal{U}|\boldsymbol{\theta}}(B_Q)$, defined by:

$$\mathtt{E}_{\mathcal{U}|\boldsymbol{\theta}}(B_Q) := \frac{\mathtt{E}_{\mathcal{U}\wedge\boldsymbol{\theta}}(B_Q)}{\pi(m_Q(\mathbf{x}',k) \geq \theta_k)},$$

where $\pi(m_Q(\mathbf{x}',k) \geq \theta_k) := \frac{1}{u}\sum_{\mathbf{x}'\in \mathrm{X}_{\mathcal{U}}} \mathbb{1}_{m_Q(\mathbf{x}',k)\geq\theta_k}$ and $k := B_Q(\mathbf{x}')$, to make a trade-off between the value of the joint Bayes error rate and the number of pseudo-labeled examples. The numerator reflects proportions of mistakes on the unlabeled set when the threshold is equal to $\boldsymbol{\theta}$, whereas the denominator computes the proportions of unlabeled observations with the margin no less than the threshold for the predicted class. One would use the bound get in Theorem 1, but two algorithmic drawbacks of the theorem have to be highlighted. First, the bound depends on the true labels of the observations. Finally, the theorem assumes that a bound for the Gibbs conditional risk is given. To avoid these issues, we take into consideration the non-deterministic case, namely, we suppose the posterior distribution $P_Y(y|\mathbf{x})$ defined over $\mathcal{Y}$. Then, we replace the deterministic $\mathbb{1}_{y=i}$ value by the corresponding probabilistic $P_Y(i|\mathbf{x})$ one. In practice, $P_Y(y|\mathbf{x})$ is approximated by $m_Q(\mathbf{x},y)$, saying that the confidence get by the margin are used as probabilities. Remark that if the space describes the problem poorly, the majority vote classifier is not able to give good margins, and then the pseudo-labelling approach can not provide a high increase in performance.

Similarly to the self-learning algorithm introduced in Amini, Laviolette, and Usunier (2008), in practice, to find an optimal θ^* we perform grid search that is the exhaustive search over the grid of values within the interval $(0,1]$. The same algorithm is used for computing the optimal γ^* that provides the value of an upper bound for the conditional risk (see Theorem 1). In contrast to the self-learning algorithm, the direct grid search in the multi-class setting is costly ($O(S^K)$, where S is the sampling rate of the grid). As

$$\mathtt{E}_{\mathcal{U}|\boldsymbol{\theta}}(B_Q) \leq \sum_{j=1}^{K} \frac{\mathtt{E}^{(j)}_{\mathcal{U}\wedge\boldsymbol{\theta}}(B_Q)}{\pi\{(m_Q(\mathbf{x}',j) \geq \theta_j) \wedge (B_Q(\mathbf{x}') = j)\}},$$

where $\mathtt{E}^{(j)}_{\mathcal{U}\wedge\boldsymbol{\theta}}(B_Q) = \sum_{i=1}^{K} \frac{u_i}{u} R_{\mathcal{U}\wedge\boldsymbol{\theta}}(B_Q,i,j)$, the sum might be minimized term by term, tuning independently each component of $\boldsymbol{\theta}$. This replaces the K-dimensional minimization task by K tasks of 1-dimensional minimization. Then, the time complexity of the threshold search is $O(K^2S^2u)$.

Algorithm 1 Multi-class self-learning algorithm (MSLA)

Input:
labeled data set $Z_{\mathcal{L}}$
Unlabeled observations $\mathrm{X}_{\mathcal{U}}$
Initialisation:
A set of pseudo-labeled instances, $Z_{\mathcal{U}} \leftarrow \emptyset$
A classifier H trained on $Z_{\mathcal{L}}$
repeat

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

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

2. $S \leftarrow \{(\mathbf{x}',y') | \mathbf{x}' \in \mathrm{X}_{\mathcal{U}}; [m_Q(\mathbf{x}',y') \geq \theta_{y'}] \wedge [y' = \mathrm{argmax}_{c\in\mathcal{Y}}\, m_Q(\mathbf{x}',c)]\}$

3. $Z_{\mathcal{U}} \leftarrow Z_{\mathcal{U}} \cup S$, $\mathrm{X}_{\mathcal{U}} \leftarrow \mathrm{X}_{\mathcal{U}} \setminus S$

4. Learn a classifier H with the following loss function:

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

until $\mathrm{X}_{\mathcal{U}}$ or S are $\emptyset$
Output: The final classifier H

Data set	# of labeled examples, $\lvert Z_{\mathcal{L}}\rvert$	# of unlabeled examples, $\lvert X_{\mathcal{U}}\rvert$	Dimension, d	# of classes, K
Vowel	99	891	10	11
DNA	31	3155	180	3
Pendigits	109	10883	16	10
MNIST	175	69825	900	10
SensIT	49	98479	100	3

Table 1: Characteristics of data sets used in our experiments ordered by the size of the training set ($|S| = |Z_{\mathcal{L}}| + |X_{\mathcal{U}}|$).

The extended strategy denoted by MSLA is described in Algorithm 1[1].

Experimental Results

In our experiments, we considered the Random Forest model with 200 trees and the maximal depth of trees (Breiman 2001), denoted by H in Algorithm 1, as the majority voted classifier with uniform posterior distribution. In this case, the margin $\mathbf{m_x}$ of an observation is evaluated by the mean vector of votes that the trees of the forest give to each class. As the size of the labeled training examples ($|Z_{\mathcal{L}}|$) is small, we did not tune the hyperparameters of the classifier and left them by their default values.

The proposed MSLA algorithm, with margin thresholds estimated by minimizing the conditional Bayes error rate, is compared with

- a supervised Random Forest (RF) trained using only labeled examples. The approach is obtained at the initialization step of MSLA and once learned it is directly applied to predict the class labels of the whole unlabeled set;
- a scikit-learn implementation of the graph based, label propagation (Pedregosa et al. 2011) approach (denoted by LP);
- the one-versus-all extension of TSVM (Joachims 1999) denoted by OVA−TSVM. In some cases, the convergence time was too long, we stopped learning the model when the convergence took more than one hour;
- the multi-class extension of the classical self-learning approach (denoted by FSLA) described in Tür, Hakkani-Tür, and Schapire (2005) with the margin thresholds fixed to the best threshold (0.7 for all classes) that we found on the unlabeled set, after testing different values manually over a predefined set of thresholds in the set $\{0.1, 0.2, \ldots, 0.9\}$.

Experiments are conducted on 5 publicly available data sets (Dheeru and Karra Taniskidou 2017; Chang and Lin 2011). The associated applications are image classification, with the MNIST and the Pendigits databases of handwritten digits; a signal processing kind of application with the SensIT data set for vehicle type classification, speech recognition using the Vowel database and finally DNA prediction using the DNA data set. We use available preprocessed versions (Chang and Lin 2011) of all data sets, except MNIST, for which we extracted HOG-features (Dalal and Triggs 2005) with the following parameters: cells of size (4, 4), blocks of size (5, 5) and the number of orientations was fixed to 4. The main characteristics of these data sets are summarized in Table 1.

Each experiment is conducted 20 times, by randomly splitting the labeled and the unlabeled training sets from the original data sets by keeping fixed their respective size (l and u) at each iteration. Results are evaluated over each unlabeled set using the accuracy (ACC) and the standard F1 measure (F1) (Baeza-Yates and Ribeiro-Neto 1999), which is the harmonic average of precision and recall. Reported performances are averaged over the 20 trials.

Table 2 summarizes results obtained by RF, FSLA, LP, OVA−TSVM and MSLA. We used bold face to indicate the highest performance rates and the symbol $\downarrow$ indicates that the performance is significantly worse than the best result, according to Mann-Whitney U test (Mann and Whitney 1947) used at the p-value threshold of 0.01. From these results it comes out that

- compared to the fully supervised approach, RF, unlabeled training data may degrade performance in some cases. This may be due to the fact that the learning hypotheses of the learning algorithms are not met regarding the data sets where the decrease is observed;
- LP and OVA−TSVM did not pass the scale over larger data sets (SensIT and MNIST).
- Self-training approaches are more robust to the large-scale problem and MSLA provides significantly better results than other approaches on Pendigits, SensIT and MNIST.
- On DNA, with a very few number of labeled training examples , OVA−TSVM outperforms MSLA.

Our analysis of these results is that the self-training algorithm does better pass the scale but it is extremely sensitive to the choice of the initial classifier and the threshold used for pseudo-labeling. On DNA the number of labeled examples is too small, leading to a bad initialization of the first classifier trained over the labeled training set. The poor estimation of the margin $m_Q(\mathbf{x}, y)$ leads to a bad approximation of the conditional probability $P_Y(y|\mathbf{x})$ used in pseudo-labeling. On SensIT, Pendigits and MNIST collections (especially the two last), the initial RF classifier is efficient, but compared to FSLA, it comes out that the choice of the threshold for pseudo-labeling is crucial, and that using the

[1]The code source of the algorithm can be found at https://github.com/vfeofanov/trans-bounds-maj-vote.

Data set	Score	RF	LP	OVA−TSVM	FSLA$_{\theta=0.7}$	MSLA
Vowel	ACC	.5832 ± .0261	.5768 ± .0268	NA	.516$^{\downarrow}$ ± .0429	**.5918** ± .0267
	F1	.5716 ± .0275	.568 ± .0261	NA	.4934$^{\downarrow}$ ± .0459	**.5804** ± .0298
DNA	ACC	.6932$^{\downarrow}$ ± .0721	.5383$^{\downarrow}$ ± .0387	**.8125** ± .0386	.5164$^{\downarrow}$ ± .0899	.7059$^{\downarrow}$ ± .0826
	F1	.65$^{\downarrow}$ ± .1086	.5348$^{\downarrow}$ ± .0437	**.8119** ± .0375	.3724$^{\downarrow}$ ± .0959	.6631$^{\downarrow}$ ± .1177
Pendigits	ACC	.8639$^{\downarrow}$ ± .022	.7767$^{\downarrow}$ ± .0515	.667$^{\downarrow}$ ± .0225	.8474$^{\downarrow}$ ± .0352	**.8866** ± .019
	F1	.8613$^{\downarrow}$ ± .0252	.7564$^{\downarrow}$ ± .0687	.6562$^{\downarrow}$ ± .0213	.8415$^{\downarrow}$ ± .0424	**.8851** ± .0198
MNIST	ACC	.8647$^{\downarrow}$ ± .0176	NA	NA	.7998$^{\downarrow}$ ± .0587	**.9085** ± .0182
	F1	.8633$^{\downarrow}$ ± .0193	NA	NA	.7743$^{\downarrow}$ ± .077	**.9086** ± .0182
SensIT	ACC	.67 ± .0291	NA	NA	.6192$^{\downarrow}$ ± .0366	**.6745** ± .0288
	F1	.654 ± .0448	NA	NA	.5784$^{\downarrow}$ ± .0683	**.6599** ± .0421

Table 2: The result table of the classification performance on different data sets described in Table 1. The performance is computed using two score functions: accuracy and F1. The sign $^{\downarrow}$ shows if the performance is statistically worse than the best result on the level 0.01 of significance. NA indicates the case when the algorithm does not converge.

conditional Bayes error rate, the margin of observations are good indicators to find such efficient thresholds. In the case of MNIST, the increase in performace compared to RF is about 4% on both the accuracy and the F1 measure.

We also analyze the behavior of the various algorithms for growing initial amounts of labeled data in the training set. Figure 1 illustrates this by showing the accuracy on a subsample of 3500 observations on MNIST of RF, FSLA$_{\theta=0.7}$ and MSLA with respect to the percentage of the labeled training examples. As expected, all performance curves increase monotonically with respect to the additional labeled data. When there are sufficient labeled training examples, all algorithms actually converge to the same accuracy performance, suggesting that the labeled data carries out sufficient information and no additional information could be extracted from unlabeled examples.

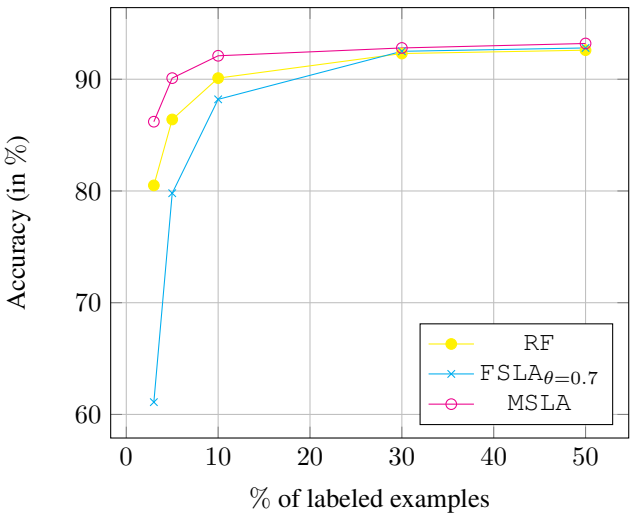

Figure 1: Classification accuracy over a subset of 3500 observations of the MNIST collection. On the graphs, dots represent the average performance on the unlabeled training set over 20 random splits.

For a low number of labeled training data, the contribution of FSLA$_{\theta=0.7}$ and MSLA that use unlabeled data is clearly shown. Even when the initial supervised RF classifier is efficient, an inexact threshold used for pseudo-labeling will lead to an addition of extra noise, making that the Random Forest trained over the augmented noisy data set becomes less effective than RF itself.

5 Conclusion

In this paper we proposed a bound over the transductive risk of a multi-class voted majority classifier. We showed how the bound can be obtained by considering the class confusion matrix as an error indicator, by involving the margin distribution of the classifier over each class and a bound over the risk of the associated Gibbs classifier. From our study, it came out that when the latter bound is tight and the errors of the majority vote classifier per class are concentrated on a low margin zone, the bound over the risk of the Bayes classifier is tight. We further showed that this bound can be solved analytically using a linear program. From this result, we then proposed to extend the self-training algorithm to the multi-class case by automatically finding a threshold for which the risk of the majority vote is the lowest. We provided empirical evidence of the extended algorithm compared to the case where the threshold is fixed manually, to a graph based semi-supervised approach and to the one-vs-all extension of TSVM. From the numerical results, it came out that the self-learning algorithm can better pass the scale but it is highly sensitive to the choice of the initial classifier trained over the labeled training set and to the threshold upon which unlabeled examples are pseudo-labeled. These results suggest that, considering the margin is effectively interesting.

6 Acknowledgments

Financial support was provided by a CNRS project PEPS and by IRS Grant (University Grenoble Alpes).

References

Amini, M.; Laviolette, F.; and Usunier, N. 2008. A transductive bound for the voted classifier with an application to semi-supervised learning. In *Advances in Neural Information Processing Systems (NIPS 21)*, 65–72.

Baeza-Yates, R. A., and Ribeiro-Neto, B. 1999. *Modern Information Retrieval*. Boston, MA, USA: Addison-Wesley Longman Publishing Co., Inc.

Breiman, L. 2001. Random forests. *Machine Learning* 45(1):5–32.

Chang, C.-C., and Lin, C.-J. 2011. LIBSVM: A Library for Support Vector Machines. *ACM Transactions on Intelligent Systems and Technology* 2(3):27:1–27:27.

Chapelle, O.; Schölkopf, B.; and Zien, A. 2010. *Semi-Supervised Learning*. The MIT Press, 1st edition.

Dalal, N., and Triggs, B. 2005. Histograms of oriented gradients for human detection. In *Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05) - Volume 1*, 886–893. Washington, DC, USA: IEEE Computer Society.

Dheeru, D., and Karra Taniskidou, E. 2017. UCI machine learning repository.

Joachims, T. 1999. Transductive inference for text classification using support vector machines. In *Proceedings of the Sixteenth International Conference on Machine Learning*, ICML '99, 200–209. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc.

Mann, H. B., and Whitney, D. R. 1947. On a test of whether one of two random variables is stochastically larger than the other. *The Annals of Mathematical Statistics* 18(1):50–60.

Maximov, Y.; Amini, M.-R.; and Harchaoui, Z. 2018. Rademacher complexity bounds for a penalized multi-class semi-supervised algorithm. *Journal of Artificial Intelligence Research* 61(1):761–786.

Morvant, E.; Koço, S.; and Ralaivola, L. 2012. PAC-Bayesian Generalization Bound on Confusion Matrix for Multi-Class Classification. In *International Conference on Machine Learning (ICML)*, 815–822. Arxiv: http://arxiv.org/abs/1202.6228, Accepted at ICML 2012.

Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; Vanderplas, J.; Passos, A.; Cournapeau, D.; Brucher, M.; Perrot, M.; and Duchesnay, E. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research* 12:2825–2830.

Tür, G.; Hakkani-Tür, D. Z.; and Schapire, R. E. 2005. Combining active and semi-supervised learning for spoken language understanding. *Speech Communication* 45:171–186.