Bornes du vote majoritaire multi-classes en présence du bruit sur les étiquettes de classes

Vasilii Feofanov*, Emilie Devijver, et Massih-Reza Amini

Université Grenoble Alpes, CNRS
Laboratoire d'Informatique de Grenoble
Bâtiment IMAG; 700 av. Centrale
38041 Saint-Martin d'Hères

Abstract

Dans ce travail, nous considérons le cadre de classification multi-classes avec des exemples d'apprentissage présentant des imperfections dans leurs étiquettes de classes. Nous modélisons cette imperfection avec un modèle d'erreur probabiliste. Sur cette base, nous dérivons des garanties théoriques pour un classifieur de vote majoritaire en étendant la borne C multi-classes, une borne supérieure du second ordre. Enfin, nous montrons empiriquement le comportement de la borne et discutons de son application pour les approches semi-supervisé basées sur le pseudo-étiquetage, en particulier pour l'auto-apprentissage.

Keywords: Noisy Labels, Ensemble Methods, Semi-supervised Learning.

1 Introduction

We consider classification learning problems where training examples are available only with imperfect labels. This is for example the case of semi-supervised learning [CSZ10], where the available set of perfectly labeled examples is scarce due to expensive data annotation, while unlabeled data are abundant. In this context, a model learned on the labeled examples only usually leads to poor learning performance, so the unlabeled examples are often incorporated to the training set along with pseudo-labels obtained through self-learning [THTS05] or co-training [BM98]. This situation makes the context different from the classical supervised setting, since the pseudo-labels may be erroneous thereby making analysis of a learning algorithm more intricate.

In this paper, we tackle this problem from a theoretical point of view for the multi-class classification case and analyze the behavior of majority vote classifiers (also known as Bayes classifiers, including Random Forest [LIS19], AdaBoost [GLL$^+$15], SVM [FTAGU15] and neural networks [LGGL19]). The majority vote classifier is well studied in the binary case, where a classical approach is to bound the majority vote risk indirectly by twice the risk of related stochastic Gibbs classifier [LST03], which, up to a linear transformation, is equivalent to the first statistical moment of the Bayes' prediction margin [GLL$^+$15]. However, the voters may compensate the errors of each other, so the majority vote risk will be much smaller than the Gibbs risk. In this connection, the majority vote's risk has been proposed to be directly upper bounded via the so-called C-bound [LLM$^+$07], which is based on the mean and the variance of the prediction margin, so it reflects both the individual strength of voters and their correlation in prediction. Nevertheless, the application of C-bound is limited by the classical supervised setting and by assuming that all training examples are perfectly labeled.

To overcome this, in our work, we take explicitly into account possible mislabeling by considering a mislabeling error model of [Chi80]. At first, we show the connection between the true and the imperfect label in misclassification of a particular example for the majority vote classifier. Then, we derive a new probabilistic C-bound over the error of the multi-class majority vote classifier in the presence of imperfect labels. Then, we derive an extension of the multi-class C-bound [LMRR17] for the probabilistic error in the presence of imperfect labels. This bound allows us to evaluate the generalization error of classification algorithms learned on mislabeled data, and particularly

*Prenom.Nom@univ-grenoble-alpes.fr

semi-supervised algorithms based on pseudo-labeling.

The rest of this paper is organized as follows. In Section 2 we introduce the problem statement and the proposed framework. Section 4 shows how to derive the C-bound in the probabilistic framework taking into account mislabeling errors. In Section 5, we illustrate the behavior of the new C-bound on real data sets. Finally, in Section 6 we summarize the outcome of this study and discuss the future work.

2 Framework and Definitions

Consider a multi-class classification problem with an input space $\mathcal{X} \subset \mathbb{R}^d$ and an output space $\mathcal{Y} = \{1, \dots, K\}$, $K \geq 2$. We denote by $\mathbf{X} = (X_1, \dots, X_d) \in \mathcal{X}$ (resp. $Y \in \mathcal{Y}$) an input (resp. output) random variable. We assume that training examples $\mathbf{x}_i \in \mathcal{X}$ are drawn *i.i.d.* according to the fixed yet unknown probability distribution $P(\mathbf{X})$, and their true labels y_i are generated according to $P(Y|\mathbf{X} = \mathbf{x}_i)$.

In this work, a fixed class of classifiers $\mathcal{H} = \{h | h : \mathcal{X} \to \mathcal{Y}\}$, called the *hypothesis space*, is considered and defined without reference to the training set. Over $\mathcal{H}$, a posterior probability distribution Q is defined after observing the training set. Further, we focus on the *Q-weighted majority vote classifier* (also called the Bayes classifier)[1] defined for all $\mathbf{x} \in \mathcal{X}$ as:

$$B_Q(\mathbf{x}) := \underset{c \in \{1,\dots,K\}}{\operatorname{argmax}} \left[\mathbb{E}_{h\sim Q} \mathbb{1}_{h(\mathbf{x})=c} \right], \qquad (1)$$

which represents a class of learning methods, where the predictions of hypotheses are aggregated using the majority vote rule scheme.

The goal of learning is formulated as to choose a posterior distribution Q over $\mathcal{H}$ based on a given training set such that the classifier B_Q will have the smallest possible error value. Compared to many works like [GLL+15, LMRR17] where the deterministic case is considered, i.e. for each example there is one and only one possible label, in this paper, we consider the more general *probabilistic* case assuming possibility of multiple outcomes for each example. Thus, we are focused on minimization of the *probabilistic risk*, which is defined as follows:

$$R(B_Q) := \mathbb{E}_{P(\mathbf{X})} \sum_{\substack{c \in \{1,\dots,K\} \\ c \neq B_Q(\mathbf{x})}} P(Y = c | \mathbf{X} = \mathbf{x}). \qquad (2)$$

To measure confidence of the majority vote classifier in its prediction, the notions of class votes and margin are further considered. Given an observation $\mathbf{x}$ and class $c \in \mathcal{Y}$, we define a class vote $v_Q(\mathbf{x}, c)$ that corresponds to the vote given by the majority vote classifier B_Q to class c for an example $\mathbf{x}$:

$$v_Q(\mathbf{x}, c) := \mathbb{E}_{h\sim Q} \mathbb{1}_{h(\mathbf{x})=c} = \sum_{h : h(\mathbf{x})=c} Q(h).$$

In practice, the vote $v_Q(\mathbf{x}, c)$ can be regarded as an estimation of the posterior probability $P(Y = c|\mathbf{X} = \mathbf{x})$; a large value indicates the high confidence of the classifier that the true label of $\mathbf{x}$ is c.
Given an observation $\mathbf{x}$, its *margin* is defined in the following way:

$$M_Q(\mathbf{x}, y) := v_Q(\mathbf{x}, y) - \max_{\substack{c \in \mathcal{Y} \\ c \neq y}} v_Q(\mathbf{x}, c). \qquad (3)$$

The margin measures a gap between the vote of the true class and the maximal vote among all other classes. If the value is strictly positive for an example $\mathbf{x}$, then y will be the output of the majority vote, so the example will be correctly classified.

3 Mislabeling Error Model

In order to model potential mislabeling of training examples, we consider an imperfect output $\hat{Y}$ that has a different distribution from the true output Y. More specifically, we summarize the label imperfection through the *mislabeling matrix* $\mathbf{P} = (p_{j,c})_{1 \leq j,c \leq K}$, defined by:

$$P(\hat{Y} = j | Y = c) := p_{j,c} \quad \forall (j, c) \in \{1, \dots, K\}^2, \qquad (4)$$

where $\sum_{j=1}^{K} p_{j,c} = 1$. We additionally assume that $\hat{Y}$ does not influence the true distribution of classes, i.e. $P(\mathbf{X}|Y, \hat{Y}) = P(\mathbf{X}|Y)$. This implies that

$$P(\hat{Y} = j | \mathbf{X} = \mathbf{x}) = \sum_{c=1}^{K} p_{j,c} P(Y = c | \mathbf{X} = \mathbf{x}). \qquad (5)$$

This class-related model is a common approach to deal with the label imperfection [Chi80, AG03, NDRT13, Sco15].

At first, we find the connection between the error of the true and the imperfect label in misclassifying a

[1] For the sake of brevity, we will tend to use the latter name, which should not be confused with other learning paradigms based on the Bayesian inference, e.g. the Bayesian statistics.

particular example $\mathbf{x} \in \mathcal{X}$. We denote

$$r(\mathbf{x}) = \sum_{\substack{c\in\{1,\dots,K\}\\ c\neq B_Q(\mathbf{x})}} P(Y = c|\mathbf{X} = \mathbf{x}),$$

$$\hat{r}(\mathbf{x}) = \sum_{\substack{c\in\{1,\dots,K\}\\ c\neq B_Q(\mathbf{x})}} P(\hat{Y} = c|\mathbf{X} = \mathbf{x}).$$

Theorem 3.1. *Let* $\mathbf{P}$ *be the mislabeling matrix, and assume that* $p_{i,i} > p_{i,j},\ \forall i, j \in \{1, \dots, K\}^2$. *Then, for all choice of* Q *on a hypothesis space* $\mathcal{H}$ *we have, for all* $\mathbf{x} \in \mathcal{X}$,

$$r(\mathbf{x}) \leq \frac{\hat{r}(\mathbf{x})}{\delta(\mathbf{x})} - \frac{1 - \alpha(\mathbf{x})}{\delta(\mathbf{x})}, \tag{6}$$

with $\delta(\mathbf{x}) := p_{B_Q(\mathbf{x}),B_Q(\mathbf{x})} - \max_{j\in\mathcal{Y}\setminus\{B_Q(\mathbf{x})\}} p_{B_Q(\mathbf{x}),j}$ *and* $\alpha(\mathbf{x}) := p_{B_Q(\mathbf{x}),B_Q(\mathbf{x})}$.

Sketch Proof. First, from the definition of $\hat{r}(\mathbf{x})$ and applying (5) we obtain that

$$\hat{r}(\mathbf{x}) = 1 - \sum_{j=1}^{K} p_{B_Q(\mathbf{x}),j} P(Y = j|\mathbf{X} = \mathbf{x})$$

Let us denote $\bar{\mathcal{Y}}_{\mathbf{x}} := \mathcal{Y} \setminus \{B_Q(\mathbf{x})\}$. Then, it can be noticed that

$$\sum_{j\in\bar{\mathcal{Y}}_{\mathbf{x}}} p_{B_Q(\mathbf{x}),j} P(Y = j|\mathbf{X} = \mathbf{x}) \leq \max_{j\in\bar{\mathcal{Y}}_{\mathbf{x}}} p_{B_Q(\mathbf{x}),j}\, r(\mathbf{x}).$$

As $p_{B_Q(\mathbf{x}),B_Q(\mathbf{x})} P(Y = B_Q(\mathbf{x})|\mathbf{X} = \mathbf{x}) = p_{B_Q(\mathbf{x}),B_Q(\mathbf{x})} - p_{B_Q(\mathbf{x}),B_Q(\mathbf{x})} r(\mathbf{x})$, we infer the following inequality:

$$\hat{r}(\mathbf{x}) \geq \delta(\mathbf{x}) r(\mathbf{x}) + 1 - \alpha(\mathbf{x}). \tag{7}$$

Taking into account the assumption that $p_{i,i} > p_{i,j},\ \forall i, j \in \{1, \dots, K\}^2$, we deduce that $\delta(\mathbf{X}) > 0$, which concludes the proof. □

This theorem gives us insights on how the true error can be bounded given the error of the imperfect label and the mislabeling matrix. With the quantities $\delta(\mathbf{x})$ and $\alpha(\mathbf{x})$, we perform a correction of $\hat{r}(\mathbf{x})$. Note that when there is no mislabeling, $\alpha(\mathbf{x}) = 1$ and $\delta(\mathbf{x}) = 1$, so the true error rate is obtained.

Note that this theorem holds also for a more general case when correction probabilities depend on the example $\mathbf{x}$. In this case, all probabilities $p_{i,j}$ are replaced by $p^{\mathbf{x}}_{i,j} := P(\hat{Y} = i|Y = j, \mathbf{X} = \mathbf{x})$. Since it is harder to estimate $p^{\mathbf{x}}_{i,j}$ compared to $p_{i,j}$, we stick to consider the class-related model described in Eq. (5).

In the theorem the mislabeling matrix is assumed given, while in practice it has to be estimated. Since the number of matrix entries grows quadratically with the increase of K, the model (5) may be more affected by the estimation error than the bound itself as the latter needs to know only $2K$ entries. We give more details about estimation of the mislabeling matrix in Section 6.

The bound can be compared with a bound derived in [Chi80, Eq. (3.14), p. 284] for the optimal Bayes classifier (maximum a-posteriori rule). It is shown that $r(\mathbf{x}) \leq 1 - \frac{1-\hat{r}(\mathbf{x})}{\beta}$, where $\beta = \max_{i=1,\dots,K}\left(\sum_{j=1}^{K} p_{i,j}\right)$. One can notice that the regularizer β is constant with respect to $\mathbf{x}$, so the penalization of the error $\hat{r}(\mathbf{x})$ does not depend on the label the classifier predicts. Another limitation is that the bound assumes that the Bayes classifier is optimal, while our bound holds for any posterior Q.

The assumption of Theorem 3.1 requires that the diagonal entries of the mislabeling matrix are the largest elements in their corresponding columns, which means that the imperfect label is reasonably correlated with the true label. However, in practice, the assumption may not hold, so the theorem is not applicable. To overcome this, the bound can be relaxed by considering $\lambda > 0$ such that $\lambda + \delta(\mathbf{x}) > 0$, so we obtain for all choices of Q on a hypothesis space $\mathcal{H}$:

$$r(\mathbf{x}) \leq \frac{\hat{r}(\mathbf{x})}{\lambda + \delta(\mathbf{x})} - \frac{1 - \lambda - \alpha(\mathbf{x})}{\lambda + \delta(\mathbf{x})}. \tag{8}$$

When $\delta(\mathbf{x})$ is close to 0, it also avoids the bound to become arbitrarily large.

4 Probabilistic C-Bound with Imperfect Labels

In this section, we derive a new risk bound in the presence of imperfect labels by combining the result obtained in Theorem 3.1 with the C-bound.

4.1 Ordinary C-Bound

[LLM$^+$07] proposed to upper bound the Bayes error by taking into account the mean and the variance of the prediction margin. A similar result was obtained in a different context by [Bre01]. [LMRR17] extended this bound to the multi-class case.

All these results were formulated in the deterministic case, but they can be further generalized to upper bound the probabilistic risk (2). In the following theorem, we present the multi-class C-bound in the probabilistic setting.

Theorem 4.1. *Let M be a random variable such that $[M|\mathbf{X}=\mathbf{x}]$ is a discrete random variable that is equal to the margin $M_Q(\mathbf{x},c)$ with probability $P(Y=c|\mathbf{X}=\mathbf{x})$, $c=\{1,\dots,K\}$. Let μ_1^M and μ_2^M be respectively the first and the second statistical moments of the random variable M. Then, for all choice of Q on a hypothesis space $\mathcal{H}$, and for all distributions $P(\mathbf{X})$ over $\mathcal{X}$ and $P(Y|\mathbf{X})$ over $\mathcal{Y}$, such that $\mu_1^M>0$, we have:*

$$R(B_Q) \leq 1 - \frac{(\mu_1^M)^2}{\mu_2^M}. \tag{CB}$$

Sketch Proof. For a fixed $\mathbf{x}$, we obtain that

$$P(M\leq 0|\mathbf{X}=\mathbf{x}) = \sum_{\substack{c\in\{1,\dots,K\}\\ c\neq B_Q(\mathbf{x})}} P(Y=c|\mathbf{X}=\mathbf{x}).$$

Applying the total probability law, we obtain that the Bayes risk is expressed as the probability of having a non-positive margin: $R(B_Q)=P(M\leq 0)$. Similarly to [LMRR17], we apply the Cantelli-Chebyshev inequality and infer the final inequality (CB). □

The main advantage of C-bound is the involvement of the second margin moment, which can be related to correlations between hypotheses' predictions, as it was shown in [LLM+07].

4.2 C-Bounds with Imperfect Labels

Theorem 4.1 assumes that all examples are perfectly labeled. Now, we consider the mislabeling error model described in Section 3. Remember that $R(B_Q)=\mathbb{E}_{\mathbf{X}}r(\mathbf{X})$. Then, by taking the expectation from the both sides of Ineq. (6), we obtain that

$$R(B_Q) = \mathbb{E}_{\mathbf{X}} r(\mathbf{X}) \leq \mathbb{E}_{\mathbf{X}}\frac{\hat{r}(\mathbf{X})}{\delta(\mathbf{X})} - \mathbb{E}_{\mathbf{X}}\frac{1-\alpha(\mathbf{X})}{\delta(\mathbf{X})}. \tag{9}$$

One can see that for every $\mathbf{x}$, $\hat{r}(\mathbf{x})$ is multiplied by a positive weight $1/\delta(\mathbf{X})>0$, so the first term of the right-hand side is a weighted generalization error of the imperfect label. To cope with this, we derive a weighted C-bound by proposing the next theorem.

Theorem 4.2. *Let $\hat{M}$ be a random variable such that $[\hat{M}|\mathbf{X}=\mathbf{x}]$ is a discrete random variable that is equal to the margin $\hat{M}_Q(\mathbf{x},i)$ with probability $P(\hat{Y}=i|\mathbf{X}=\mathbf{x})$, $i=\{1,\dots,K\}$. Assume that every diagonal entry of the mislabeling matrix $\mathbf{P}$ is the largest element in the corresponding column, i.e. $p_{i,i}>p_{i,j}$, $\forall i,j\in\{1,\dots,K\}^2$. Then, for all choice of Q on a hypothesis space $\mathcal{H}$, and for all distributions $P(\mathbf{X})$ over $\mathcal{X}$ and $P(Y|\mathbf{X})$ over $\mathcal{Y}$, we have:*

$$R(B_Q) \leq \psi_{\mathbf{P}} - \frac{\left(\mu_1^{\hat{M},\mathbf{P}}\right)^2}{\mu_2^{\hat{M},\mathbf{P}}}, \tag{CBIL}$$

if $\mu_1^{\hat{M}_{\mathbf{P}}}>0$, where

- *$\psi_{\mathbf{P}} := \mathbb{E}_{\mathbf{X}}\frac{\alpha(\mathbf{X})}{\delta(\mathbf{X})}$ with δ and α as in Theorem 3.1,*
- *$\mu_1^{\hat{M},\mathbf{P}} := \int_{\mathbb{R}^{d+1}}\frac{m}{\delta(\mathbf{x})}P(\hat{M}=m,\mathbf{X}=\mathbf{x})\mathrm{d}\mathbf{x}\mathrm{d}m$ is the weighted 1st margin moment,*
- *$\mu_2^{\hat{M},\mathbf{P}} := \int_{\mathbb{R}^{d+1}}\frac{m^2}{\delta(\mathbf{x})}P(\hat{M}=m,\mathbf{X}=\mathbf{x})\mathrm{d}\mathbf{x}\mathrm{d}m$ is the weighted 2nd margin moment.*

Proof. At first, let us introduce a normalization factor $\omega_{\mathbf{P}}$ defined as follows:

$$\omega_{\mathbf{P}} := \mathbb{E}_{\mathbf{X}}\frac{1}{\delta(\mathbf{X})} = \int_{\mathbb{R}^{d+1}}\frac{P(\hat{M}=m,\mathbf{X}=\mathbf{x})}{\delta(\mathbf{x})}\mathrm{d}\mathbf{x}\mathrm{d}m.$$

Remind that $\hat{r}(\mathbf{x})=P(\hat{M}\leq 0|\mathbf{X}=\mathbf{x})$. Then, we can write:

$$\begin{aligned}\mathbb{E}_{\mathbf{X}}\frac{\hat{r}(\mathbf{X})}{\delta(\mathbf{X})} &= \int_{\mathbb{R}^d}\frac{1}{\delta(\mathbf{x})}P(\hat{M}\leq 0|\mathbf{X}=\mathbf{x})P(\mathbf{X}=\mathbf{x})\mathrm{d}\mathbf{x}\\ &= \omega_{\mathbf{P}}\int_{-\infty}^{0}\frac{\int_{\mathbb{R}^d}P(\hat{M}=m,\mathbf{X}=\mathbf{x})/\delta(\mathbf{x})\mathrm{d}\mathbf{x}}{\int_{\mathbb{R}^{d+1}}P(\hat{M}=m,\mathbf{X}=\mathbf{x})/\delta(\mathbf{x})\mathrm{d}\mathbf{x}\mathrm{d}m}\mathrm{d}m.\end{aligned}$$

The expression inside the integral in the last equality is a density, which we denote it by f_ω and the corresponding random variable by $\hat{M}_\omega$. Then, we obtain that $\mathbb{E}_{\mathbf{X}}\frac{\hat{r}(\mathbf{X})}{\delta(\mathbf{X})}=\omega_{\mathbf{P}}P(\hat{M}_\omega<0)$.

We further notice that the weighted first and second moments can be represented respectively as $\mu_1^{\hat{M},\mathbf{P}}=\omega_{\mathbf{P}}\mu_1^{\hat{M}_\omega}$ and $\mu_2^{\hat{M},\mathbf{P}}=\omega_{\mathbf{P}}\mu_2^{\hat{M}_\omega}$. Also, we have $var(M_\omega)=\left(\mu_2^{\hat{M},\mathbf{P}}/\omega_{\mathbf{P}}\right)-\left(\mu_1^{\hat{M},\mathbf{P}}/\omega_{\mathbf{P}}\right)^2$. Then, using the Cantelli-Chebyshev inequality we deduce:

$$P(\hat{M}_\omega<0) \leq 1 - \frac{\left(\mu_1^{\hat{M},\mathbf{P}}\right)^2}{\omega_{\mathbf{P}}\mu_2^{\hat{M},\mathbf{P}}}. \tag{10}$$

Combining Eq. (10) and Eq. (9) we infer (CBIL):

$$R(B_Q) \leq \mathbb{E}_{\mathbf{X}}\frac{\hat{r}(\mathbf{x})}{\delta(\mathbf{x})} - \mathbb{E}_{\mathbf{X}}\frac{1-\alpha(\mathbf{x})}{\delta(\mathbf{x})} \leq \psi_{\mathbf{P}} - \frac{\left(\mu_1^{\hat{M},\mathbf{P}}\right)^2}{\mu_2^{\hat{M},\mathbf{P}}}.$$

□

Given data with imperfect labels, the direct evaluation of the generalization error rate may be biased, leading to an overly optimistic evaluation. Using the mislabeling matrix $\mathbf{P}$ we derive a more conservative C-bound, where the error of $\mathbf{x}$ is penalized by the factor $1/\delta(\mathbf{x})$. When there is no mislabeling, $\psi_{\mathbf{P}} = 1$, $\mu_1^{\hat{M},\mathbf{P}}$ and $\mu_2^{\hat{M},\mathbf{P}}$ are equivalent to $\mu_1^{\hat{M}}$ and $\mu_2^{\hat{M}}$, so we obtain the regular C-bound (CB).

Note that empirical estimation of the margin mean, the margin variance and the mislabeling matrix may induce some optimism in the bound evaluation. To overcome this, the problem can be further analyzed in the PAC-Bayesian framework initiated by [McA99] in order to derive a Probably Approximately Correct bound on $R(B_Q)$ based on the sample estimates of $\psi_{\mathbf{P}}, \mu_1^{\hat{M},\mathbf{P}}, \mu_2^{\hat{M},\mathbf{P}}$. A PAC-Bayesian bound additionally penalizes the bound by the sample size and the divergence between the posterior Q and a fixed prior distribution P defined over $\mathcal{H}$ before observing training data. Due to the lack of space, we omit derivations, but the bound can be straightforwardly obtained using the results of [LMRR17] and [Mau04].

4.3 Application to Semi-supervised Learning

In the semi-supervised setting, it is assumed available a training set of labeled examples $\{(\mathbf{x}_i, y_i)\}_{i=1}^{l}$ and unlabeled examples $\{\mathbf{x}_i\}_{i=l+1}^{l+u}$ where $l \ll u$. In order to exploit the unlabeled examples, a common approach is pseudo-label them, so we end up with an imperfectly labeled data set $\{\mathbf{x}_i, \hat{y}_i\}_{i=1}^{l+u}$. For example, a self-learning algorithm [THTS05] learns an initial classifier on the labeled training data, and then selects iteratively a subset of unlabeled examples with confidence score above a fixed threshold and include them along with their pseudo-labels in the training set to retrain the supervised classifier. [FDA19] have proposed to find this threshold dynamically as a trade-off between the number of pseudo-labeled examples and the transductive error they induce. In order to evaluate the error, they have derived a transductive bound of the majority vote error in the multi-class case.

Although the proposed by [FDA19] strategy allows to minimize the error induced by self-learning, in the end, the algorithm may be still learned on erroneous labels, and we do not know how to evaluate the classifier's error in this noisy case. In addition, their transductive bound can be regarded as a first-order bound, since it is linearly dependent on the classifier' votes, so it does not take into account the correlation between hypotheses. These two issues can be overcome by applying our result obtained in Section 4.2 to this setting.

Data set	# of lab. examples, l	# of unlab. examples, u	# of feat., d	# of classes, K
`Isolet`	389	7408	617	26
`HAR`	102	10197	561	6
`Letter`	400	19600	16	26
`MNIST`	175	69825	900	10

Table 1: Characteristics of data sets used in our experiments ordered by the size of the training set ($n = l+u$).

Thus, the (CBIL) represents a semi-supervised bound on the risk $R(B_Q)$, where B_Q is learned on the labeled examples, while the bound is evaluated on the unlabeled set pseudo-labeled by the self-learning algorithm. Comparing with the transductive bound of [FDA19], (CBIL) bounds the risk directly and not from the conditional risk, so it will be tighter in most of cases.

Note that there exists other attempts to evaluate the C-bound in the semi-supervised setting. In the binary case, [LLM$^+$07] estimated the second margin moment using additionally unlabeled data by expressing it via disagreement of hypotheses. However, this holds for the binary case only.

5 Empirical Illustration

In this section, we empirically illustrate the value of (CBIL) in the semi-supervised setting. Experiments are conducted on publicly available data sets [DG17, CL11]. In order to emulate the semi-supervised context, we do not use the train/test splits that are proposed by data sources. Instead, we propose our own splits so that $l \ll u$. Our experiment is conducted 20 times, by randomly splitting an original data set on a labeled and an unlabeled parts keeping fixed their respective size at each iteration. The reported results are averaged over the 20 trials.

Experiments are conducted on 4 real data sets. The associated applications are image classification with the `MNIST` databases of handwritten digits; a signal processing application with the human activity recognition `HAR` database; speech recognition using the `Isolet` and the `Letter` data sets. The main characteristics of these data sets are summarized in Table 1.

As a semi-supervised classifier, we take the self-learning algorithm with dynamic thresholding proposed by [FDA19]. We take the Random Forest algorithm [Bre01] with 200 trees and the maximal depth

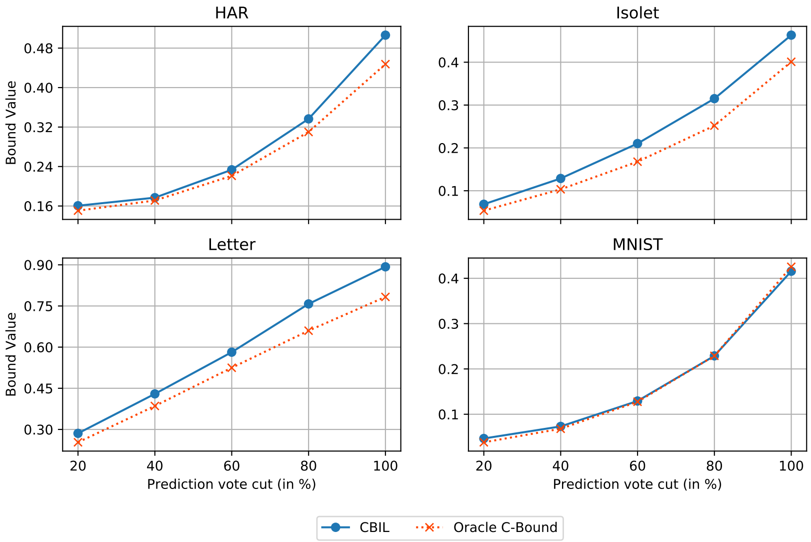

Figure 1: (CBIL) and Oracle C-Bound when varying the number of pseudo-labels on 4 data sets. We keep the most confident one (with respect to prediction vote) from 20% to 100%.

of trees as the majority vote classifier with the uniform posterior distribution. We use it also as the base classifier for self-learning. For an observation $\mathbf{x}$, we evaluate the vector of class votes $\{v(\mathbf{x}, i)\}_{i=1}^{K}$ by averaging over the trees the vote given to each class by the tree. A tree computes a class vote as the fraction of training examples in a leaf belonging to a class.

Our experiment consists in empirical study of how the value of (CBIL) evaluated on the pseudo-labeled unlabeled examples is penalized by the mislabeling model. For this, we empirically compare it with the oracle C-bound (CB) evaluated as if the labels for the considered unlabeled data would be known.

To do so, we compute the value of the two bounds varying the number of examples used for evaluation with respect to the prediction confidence: the pseudo-labeled examples are sorted by the value of the prediction vote of B_Q in the descending order, and we keep only the first $\rho\%$ of the examples for $\rho \in \{20, 40, 60, 80, 100\}$.

By using the prediction vote of B_Q we expect that with increase of ρ we have more mislabels, so the (CBIL) is more penalized. In (CBIL), we use the true value of the mislabeling matrix (i.e. evaluated using the labels of unlabeled data) for clear illustration of the C-bound's penalization. In Section 6, we discuss the possible estimations of the mislabeling matrix.

The experimental results on 4 data sets `HAR`, `Isolet`, `Letter` and `MNIST` are illustrated in Figure 1. As expected, the classifier makes mistakes mostly on low class votes, so the error increases when ρ grows. One can see that on `Isolet`, `HAR` and `Letter` (CBIL) is close to the oracle C-bound for small ρ, since most of pseudo-labels are true. When more noisy pseudo-labels are included, the difference between the two values becomes more evident, leading (CBIL) to be more pessimistic. This is probably connected with the choice of the mislabeling error model (4) that is class-related and not instance-related. Although we lose some flexibility, the class-related mislabeling matrix would be easier to estimate in practice. Finally, for `MNIST`, the two bounds are very close to each other, and the mislabeling is occasional, which is agreed with the performance of the self-learning on this data set [FDA19, p. 3572] as pseudo-labels are very helpful in this case.

6 Conclusion and Future Work

In this paper, we proposed a new probabilistic framework for analysis of the multi-class majority vote classifier in the presence of imperfect labels. We proposed a mislabeling error model to take explicitly into account these mislabeling errors and established the connection between the true and the imperfect output. Based on this result, we extended the C-bound to the case when imperfect labels are used for evaluation. The proposed bound allows us to evaluate the performance of majority vote learning models in this noisy case. In particular, the result can be applied to semi-supervised learning to deal with self-learning approaches that pseudo-label unlabeled examples.

In the semi-supervised setting, we illustrated the influence of the mislabeling error model on the bound's value on several real data sets. However, further application of C-bound raises several open practical questions for us, which we detail below and leave as a subject for future work.

Firstly, the analysis of the learning model learned on pseudo-labels is perplexing due to the so-called *confirmation bias*: at every iteration, the self-learning includes into the training set unlabeled examples with highly confident predictions, which arise from classifier's overconfidence to its initial decisions that could be erroneous. This implies that the hypotheses will have small disagreement on the unlabeled set after pseudo-labeling, so the votes are no more adequate for measuring prediction confidence. A correct estimation of mislabeling probabilities or changing the way self-learning is learned are possible solutions.

Secondly, (CBIL) requires in practice the estimation of the mislabeling matrix, which is a complex problem, but an active subject of study [NDRT13]. Most of these studies tackle this problem from an algorithmic point of view: for example, in the semi-supervised setting, [KARG08] learn the mislabeling matrix together with the classifier parameters through the classifier likelihood maximization for document classification; in the supervised setting, a common approach is to detect anchor points whose labels are surely true [Sco15]. A potential idea would be to transfer this idea to the semi-supervised case in order to detect the anchor points in the unlabeled set and use them together with the labeled set for correct estimation of the noise in pseudo-labels; this may require additional assumptions such as the existence of clusters [Rig07, MAH18] or manifold structure [BN04].

We also point out possible applications of (CBIL). At first, the bound can be used for model selection tasks as semi-supervised feature selection [SSGC17]. Since minimization of the C-bound implies simultaneously margin mean maximization and margin variance minimization, (CBIL) would guide a feature selection algorithm to choose an optimal feature subset based on the labeled and the pseudo-labeled sets.

Next, (CBIL) can be used as a criterion to learn the posterior Q in the semi-supervised setting. This issue is actively studied in the supervised context, e.g. [RML16, BCRL20] have been developed the boosting-based C-bound optimization algorithms.

It should be noticed that for these two applications, the main objective is to rank models, so the best model has the minimal error on the unlabeled set. Hence, the bound analysis goes beyond the classical question of tightness: the tightest bound does not always imply the minimal error, and a bound relaxation can have a positive effect as it is the case for Ineq. (8).

References

[AG03] Massih-Reza Amini and Patrick Gallinari. Semi-supervised learning with explicit misclassification modeling. In Georg Gottlob and Toby Walsh, editors, *IJCAI-03, Proceedings of the Eighteenth International Joint Conference on Artificial Intelligence, Acapulco, Mexico, August 9-15, 2003*, pages 555–560. Morgan Kaufmann, 2003.

[BCRL20] Baptiste Bauvin, Cécile Capponi, Jean-Francis Roy, and François Laviolette. Fast greedy c-bound minimization with guarantees. *Machine Learning*, 109(9):1945–1986, 2020.

[BM98] Avrim Blum and Tom Mitchell. Combining labeled and unlabeled data with co-training. In *Proceedings of the eleventh annual conference on Computational learning theory (COLT)*, pages 92–100, 1998.

[BN04] Mikhail Belkin and Partha Niyogi. Semi-supervised learning on riemannian manifolds. *Machine Learning*, 56(1-3):209–239, 2004.

[Bre01] Leo Breiman. Random forests. *Machine Learning*, 45(1):5–32, Oct 2001.

[Chi80] CB Chittineni. Learning with imperfectly labeled patterns. *Pattern Recognition*, 12(5):281–291, 1980.

[CL11] Chih-Chung Chang and Chih-Jen Lin. LIBSVM: A Library for Support Vector Machines. *ACM Transactions on Intelligent Systems and Technology*, 2(3):27:1–27:27, May 2011.

[CSZ10] Olivier Chapelle, Bernhard Schölkopf, and Alexander Zien. *Semi-Supervised Learning*. The MIT Press, 1st edition, 2010.

[DG17] Dheeru Dua and Casey Graff. UCI machine learning repository, 2017.

[FDA19] Vasilii Feofanov, Emilie Devijver, and Massih-Reza Amini. Transductive bounds for the multi-class majority vote classifier. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33(01):3566–3573, Jul. 2019.

[FTAGU15] Ali Fakeri-Tabrizi, Massih-Reza Amini, Cyril Goutte, and Nicolas Usunier. Multiview self-learning. *Neurocomputing*, 155(C):117–127, May 2015.

[GLL$^+$15] Pascal Germain, Alexandre Lacasse, François Laviolette, Mario March, and Jean-Francis Roy. Risk bounds for the majority vote: From a pac-bayesian analysis to a learning algorithm. *Journal of Machine Learning Research*, 16(26):787–860, 2015.

[KARG08] Anastasia Krithara, Massih-Reza Amini, Jean-Michel Renders, and Cyril Goutte. Semi-supervised document classification with a mislabeling error model. In *Advances in Information Retrieval , 30th European Conference on IR Research, ECIR 2008, Glasgow, UK, March 30-April 3, 2008. Proceedings*, pages 370–381, 2008.

[LGGL19] Gaël Letarte, Pascal Germain, Benjamin Guedj, and François Laviolette. Dichotomize and generalize: Pac-bayesian binary activated deep neural networks. In *Advances in Neural Information Processing Systems*, pages 6872–6882, 2019.

[LIS19] Stephan S Lorenzen, Christian Igel, and Yevgeny Seldin. On pac-bayesian bounds for random forests. *Machine Learning*, 108(8-9):1503–1522, 2019.

[LLM$^+$07] Alexandre Lacasse, François Laviolette, Mario Marchand, Pascal Germain, and Nicolas Usunier. Pac-bayes bounds for the risk of the majority vote and the variance of the gibbs classifier. In *Advances in Neural Information Processing Systems*, pages 769–776, 2007.

[LMRR17] François Laviolette, Emilie Morvant, Liva Ralaivola, and Jean-Francis Roy. Risk upper bounds for general ensemble methods with an application to multiclass classification. *Neurocomputing*, 219:15–25, 2017.

[LST03] John Langford and John Shawe-Taylor. Pac-bayes & margins. In *Advances in Neural Information Processing Systems*, pages 439–446, 2003.

[MAH18] Yury Maximov, Massih-Reza Amini, and Zaid Harchaoui. Rademacher complexity bounds for a penalized multi-class semi-supervised algorithm. *Journal of Artificial Intelligence Research*, 61(1):761–786, 2018.

[Mau04] Andreas Maurer. A note on the pac bayesian theorem, 2004.

[McA99] David A McAllester. Some pac-bayesian theorems. *Machine Learning*, 37(3):355–363, 1999.

[NDRT13] Nagarajan Natarajan, Inderjit S Dhillon, Pradeep K Ravikumar, and Ambuj Tewari. Learning with noisy labels. In *Advances in Neural Information Processing Systems*, pages 1196–1204, 2013.

[Rig07] Philippe Rigollet. Generalization error bounds in semi-supervised classification under the cluster assumption. *Journal of Machine Learning Research*, 8(Jul):1369–1392, 2007.

[RML16] Jean-Francis Roy, Mario Marchand, and François Laviolette. A column generation bound minimization approach with pac-bayesian generalization guarantees. In *Artificial Intelligence and Statistics*, pages 1241–1249, 2016.

[Sco15] Clayton Scott. A rate of convergence for mixture proportion estimation, with application to learning from noisy labels. In *Artificial Intelligence and Statistics*, pages 838–846, 2015.

[SSGC17] Razieh Sheikhpour, Mehdi Agha Sarram, Sajjad Gharaghani, and Mohammad Ali Zare Chahooki. A survey on semi-supervised feature selection methods. *Pattern Recognition*, 64(C):141–158, April 2017.

[THTS05] Gökhan Tür, Dilek Z. Hakkani-Tür, and Robert E. Schapire. Combining active and semi-supervised learning for spoken language understanding. *Speech Communication*, 45:171–186, 2005.