Multiview Boosting by Controlling the Diversity and the Accuracy of View-specific Voters
Anil Goyal, Emilie Morvant, Pascal Germain, Massih-Reza Amini

To cite this version:

HAL Id: hal-01857463
https://hal.archives-ouvertes.fr/hal-01857463v4
Submitted on 26 May 2019

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.
Multiview Boosting by Controlling the Diversity and the Accuracy of View-specific Voters

Anil Goyal$^{1,2}$ Emilie Morvant$^1$ Pascal Germain$^3$ Massih-Reza Amini$^2$

$^1$ Univ Lyon, UJM-Saint-Etienne, CNRS, Institut d’Optique Graduate School, Laboratoire Hubert Curien UMR 5516, F-42023, Saint-Etienne, France
$^2$ Univ. Grenoble Alps, Laboratoire d’Informatique de Grenoble, AMA, Centre Equation 4, BP 53, F-38041 Grenoble Cedex 9, France
$^3$ Inria Lille - Nord Europe, Modal Project-Team, 59650 Villeneuve d’Ascq, France

May 26, 2019

Abstract

In this paper we propose a boosting based multiview learning algorithm, referred as **PB-MVBoost**, which iteratively learns i) weights over view-specific voters capturing view-specific information, and ii) weights over views by optimizing a PAC-Bayes multiview C-Bound that takes into account the accuracy of view-specific classifiers and the diversity between the views. We derive a generalization bound for this strategy following the PAC-Bayes theory which is a suitable tool to deal with models expressed as weighted combination over a set of voters. Different experiments on three publicly available datasets show the efficiency of the proposed approach with respect to state-of-art models.

1 Introduction

With the tremendous generation of data, there are more and more situations where observations are described by more than one view. This is for example the case with multilingual documents that convey the same information in different languages or images that are naturally described according to different set of features (for example SIFT, HOG, CNN, etc). In this paper, we study the related machine learning problem that consists in finding an efficient classification model from different information sources that describe the observations. This topic, called multiview (or multimodal) learning [Atrey et al. 2010, Sun 2013, Baltrušaitis et al. 2019], has been expanding over the past decade, spurred by the seminal work of Blum and Mitchell 1998 on co-training (with only two views). The aim is to learn a classifier which performs better than classifiers trained over each view separately (called here view-specific classifier). Usually, this is done by directly concatenating the representations (early fusion) or by combining the predictions of view-specific classifiers (late fusion) Snoek et al. 2005. In this work, we stand in the latter situation. Concretely, we study a two-level multiview learning strategy based on the PAC-Bayesian theory (introduced by McAllester 1999 for monoview learning). This theory provides Probably Approximately Correct (PAC) generalization guarantees for models expressed as a weighted combination over a set of functions/voters (i.e., for a weighted majority vote). In this framework, given a prior distribution over a set of functions, called voters, $\mathcal{H}$ and a learning sample, one aims at learning a posterior distribution over $\mathcal{H}$ leading to a well-performing majority vote; each voter from $\mathcal{H}$ is weighted by its probability to appear according to the posterior distribution. Note that, PAC-Bayesian studies have not only been conducted to characterize the error of such weighted majority votes Catoni 2007, Seeger 2002, Langford and Shawe-Taylor 2002, Germain et al. 2009, 2015, but have also been used to derive theoretically grounded learning algorithms (such as for supervised learning Germain et al. 2009, Parrado-Hernández et al. 2012, Alquier et al. 2015, Roy et al. 2016, Morvant et al. 2014 or transfer learning Germain et al. 2016). To tackle multiview learning in a PAC-Bayesian fashion, we propose to define a two-level hierarchy of prior and posterior distributions over the views: i) for each view $v$, we consider a prior $P_v$ and a posterior $Q_v$ distributions over view-specific voters to capture view-specific information and ii) a hyper-prior $\pi_v$ and a hyper-posterior $\rho_v$ distributions over the set of views to capture the accuracy of view-specific classifiers and diversity.
between the views (see Figure 1). Following this distributions’ hierarchy, we define a multiview majority vote classifier where the view-specific classifiers are weighted according to posterior and hyper-posterior distributions. By doing so, we extend the classical PAC-Bayesian theory to multiview learning with more than two views and derive a PAC-Bayesian generalization bound for our multiview majority vote classifier.

From a practical point of view, we design an algorithm based on the idea of boosting [Freund 1995], Freund and Schapire [1997], Schapire [1999, 2003], an ensemble method well known to be able to learn well-performing majority vote. Our boosting-based multiview learning algorithm, called PB-MVBoost, deals with the two-level hierarchical learning strategy. PB-MVBoost is then an ensemble method that outputs a multiview classifier expressed as a weighted combination of view-specific voters. It is important to notice that controlling the diversity between the view-specific classifiers or the views is a key element in multiview learning [Amini et al. 2009], Goyal et al. [2017], Chapelle et al. [2010], Kuncheva [2004], Maillard and Vayatis [2009], Morvant et al. [2014]. Therefore, to learn the weights over the views, we minimize an upper-bound on the error of the majority vote, called the multiview C-bound [Germain et al. 2015], Roy et al. [2016], Goyal et al. [2017], allowing us to control a trade-off between accuracy and diversity. Concretely, at each iteration of our multiview algorithm, we learn i) weights over view-specific voters based on their ability to deal with examples on the corresponding view (capturing view-specific information), and ii) weights over views by minimizing the multiview C-bound. To show the potential of our algorithm, we empirically evaluate our approach on MNIST, MNIST and Reuters RCV1/RCV2 collections [Lecun et al. 1998], Amini et al. [2009]. We observe that our algorithm PB-MVBoost, empirically minimizes the multiview C-Bound over iterations, and leads to good performances even when the classes are unbalanced. We compare PB-MVBoost with a previously developed multiview algorithm, denoted by FusionCq, Goyal et al. [2017], which first learns the view-specific voters at the base level of the hierarchy, and then, combines the predictions of view-specific voters using a PAC-Bayesian algorithm CqBoost Roy et al. [2016]. From the experimental results, it came out that PB-MVBoost is more stable across different datasets and computationally faster than FusionCq.

In the next section, we discuss some related works. In Section 3 we present the PAC-Bayesian multiview learning framework Goyal et al. [2017]. In Section 4 we derive our multiview learning algorithm PB-MVBoost. Before concluding in Section 6 we experiment our algorithm in Section 5.

2 Related Work

Learning a weighted majority vote is closely related to ensemble methods [Dietterich 2000], Re and Valentini [2012]. In the ensemble methods literature, it is well known that we desire to combine voters that make errors on different data points Kuncheva [2004]. Intuitively, this means that the voters disagree on some data points. This notion of disagreement (or agreement) is sometimes called diversity between
classifiers [Diez-Pastor et al. 2015, Brown and Kuncheva 2010, Kuncheva 2004]. Even if there is no consensus on the definition of “diversity”, controlling it while keeping good accuracy is at the heart of a majority of ensemble methods: indeed if all the voters agree on all the points then there is no interest to combine them, only one will be sufficient. Similarly, when we combine multiple views (or representations), it is known that controlling diversity between the views plays a vital role for learning the final majority vote [Amini et al. 2009, Goyal et al. 2017, Chapelle et al. 2010, Maillard and Vayatis 2009]. Most of the existing ensemble-based multiview learning algorithms try to exploit either view consistency (agreement between views) Janodet et al. [2009], Koç and Capponi [2011], Xiao and Guo [2012] or diversity between views Xu and Sun [2010], Goyal et al. [2017], Peng et al. [2011, 2017] in different manners. Janodet et al. [2009] proposed a boosting based multiview learning algorithm for two views, called 2-Boost. At each iteration, the algorithm learns the weights over the view-specific voters by maintaining a single distribution over the learning examples. Conversely, Koç and Capponi [2011] proposed Munbo that maintains separate distributions for each view. For each view, the algorithm reduces the weights associated with the examples hard to classify, and increases the weights of those examples in the other views. This trick allows a communication between the views with the objective to maintain view consistency. Compared to our approach, we follow a two-level learning strategy where we learn (hyper-)posterior distributions/weights over view-specific voters and views. In order to take into account accuracy and diversity between the views, we optimize the multiview C-Bound (an upper-bound over the risk of multiview majority vote learned, see e.g. Germain et al. [2015], Roy et al. [2016], Goyal et al. [2017]).

Xu and Sun [2010] proposed EMV-AdaBoost, an embedded multiview AdaBoost algorithm, restricted to two views. At each iteration, an example contributes to the error if it is misclassified by any of the view-specific voters and the diversity between the views is captured by weighting the error by the agreement between the views. Peng et al. [2011, 2017] proposed variants of Boost.SH (boosting with SHared weight distribution) which controls the diversity for more than two views. Similarly than our approach, they maintain a single global distribution over the learning examples for all the views. To control the diversity between the views, at each iteration they update the distribution over the views by casting the algorithm in two ways: i) a multiarmed bandit framework (rBoost.SH) and ii) an expert strategy framework (eBoost.SH) consisting of set of strategies (distribution over views) for weighing views. At the end, their multiview majority vote is a combination of T weighted base voters, where T is the number of iterations for boosting. Whereas, our multiview majority vote is a weighted combination of the view-specific voters over all the weighted views.

Furthermore, our approach encompasses the one of Amini et al. [2009] and Xiao and Guo [2012], Amini et al. [2009] proposed a Rademacher analysis for expectation of individual risks of each view-specific classifier (for more than two views). Xiao and Guo [2012] derived a weighted majority voting Adaboost algorithm which learns weights over view-specific voters at each iteration of the algorithm. Both of these approaches maintain a uniform distribution over the views whereas our algorithm learns the weights over the views such that they capture diversity between the views. Moreover, it is important to note that Sun et al. [2017] proposed a PAC-Bayesian analysis for multiview learning over the concatenation of views but limited to two views and to a particular kind of voters: linear classifiers. This has allowed them to derive a SVM-like learning algorithm but dedicated to multiview with exactly two views. In our work, we are interested in learning from more than two views and without any restrictions on the classifier type. Contrary to them, we followed a two-level distributions’ hierarchy where we learn weights over view-specific classifiers and weights over views.

3 The Multiview PAC-Bayesian Framework

3.1 Notations and Setting

In this work, we tackle multiview binary classification tasks where the observations are described with \( V \geq 2 \) different representation spaces, i.e., views. Let \( \mathcal{V} \) be the set of these \( V \) views. Formally, we focus on tasks for which the input space is \( \mathcal{X} = \mathcal{X}_1 \times \cdots \times \mathcal{X}_V \), where \( \forall v \in \mathcal{V}, \mathcal{X}_v \subseteq \mathbb{R}^{d_v} \) is a \( d_v \)-dimensional input space, and the binary output space is \( \mathcal{Y} = \{-1, +1\} \). We assume that \( \mathcal{D} \) is a fixed but unknown distribution over \( \mathcal{X} \times \mathcal{Y} \). We stand in the PAC-Bayesian supervised learning setting where an observation \( x = (x^1, x^2, \ldots, x^V) \in \mathcal{X} \) is given with its label \( y \in \mathcal{Y} \), and is independently and identically drawn (i.i.d.) from \( \mathcal{D} \). A learning algorithm is then provided with a training sample \( S \) of \( n \) examples i.i.d. from \( \mathcal{D} \): \( S = \{(x_i, y_i)\}_{i=1}^n \sim (\mathcal{D})^n \), where \((\mathcal{D})^n \) stands for the distribution of a \( n \)-sample. For each view \( v \in \mathcal{V} \), we consider a view-specific set \( \mathcal{H}_v \) of voters \( h : \mathcal{X}_v \to \mathcal{Y} \), and a prior distribution \( P_0 \) on \( \mathcal{H}_v \). Given a hyper-prior distribution \( \pi \) over the views \( \mathcal{V} \), and a multiview learning sample \( S \), our PAC-Bayesian learner...
objective is twofold: i) finding a posterior distribution \(Q_v\) over \(\mathcal{H}_v\) for all views \(v \in \mathcal{V}\), and ii) finding a hyper-posterior distribution \(\rho\) on the set of the views \(\mathcal{V}\). This defines a hierarchy of distributions illustrated on Figure 1. The learned distributions express a multiview weighted majority vote\(^1\) defined as

\[
B_\rho(x) = \text{sign} \left[ \mathbb{E}_{v \sim \rho} \mathbb{E}_{h \sim Q_v} h(x^v) \right].
\]

Thus, the learner aims at constructing the posterior and hyper-posterior distributions that minimize the true risk \(R_D(B_\rho)\) of the multiview weighted majority vote

\[
R_D(B_\rho) = \mathbb{E}_{(x,y) \sim D} \mathbb{I}_{[B_\rho(x) \neq y]},
\]

where \(\mathbb{I}[\pi] = 1\) if the predicate \(\pi\) is true and 0 otherwise. The above risk of the deterministic weighted majority vote is closely related to the Gibbs risk \(R_D(G_\rho)\) defined as the expectation of the individual risks of each voter that appears in the majority vote. More formally, in our multiview setting, we have

\[
R_D(G_\rho) = \mathbb{E}_{(x,y) \sim D} \mathbb{E}_{v \sim \rho} \mathbb{E}_{h \sim Q_v} \mathbb{I}_{[h(x^v) \neq y]},
\]

and its empirical counterpart is

\[
R_S(G_\rho) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}_{v \sim \rho} \mathbb{E}_{h \sim Q_v} \mathbb{I}_{[h(x^v_i) \neq y_i]}.
\]

In fact, if \(B_\rho\) misclassifies \(x \in X\), then at least half of the view-specific voters from all the views (according to hyper-posterior and posterior distributions) makes an error on \(x\). Then, it is well known \cite{Shawe-Taylor_2002, Langford_2003, McAllester_2003, Germain_2015} that \(R_D(B_\rho)\) is upper-bounded by twice \(R_D(G_\rho)\):

\[
R_D(B_\rho) \leq 2R_D(G_\rho).
\]

In consequence, a generalization bound for \(R_D(G_\rho)\) gives rise to a generalization bound for \(R_D(B_\rho)\).

There exist tighter relations \cite{Langford_2002, Germain_2015, Lacasse_2006}, such as the C-Bound \cite{Lacasse_2006, Germain_2015} which captures a trade-off between the Gibbs risk \(R_D(G_\rho)\) and the disagreement between pairs of voters. This latter can be seen as a measure of diversity among the voters involved in the majority vote \cite{Roy_2011, Morvant_2014}, that is a key element to control from a multiview point of view \cite{Atrey_2010, Amini_2009, Goyal_2017, Kuncheva_2004, Maillard_2009}. The C-Bound can be extended to our multiview setting as below.

**Lemma 1 (Multiview C-Bound)** Let \(V \geq 2\) be the number of views. For all posterior \(\{Q_v\}_{v=1}^{V}\) distributions over \(\{\mathcal{H}_v\}_{v=1}^{V}\) and hyper-posterior \(\rho\) distribution over views \(V\), if \(R_D(G_\rho) < \frac{1}{2}\), then we have

\[
R_D(B_\rho) \leq 1 - \frac{(1 - 2R_D(G_\rho))^2}{1 - 2d_D(\rho)},
\]

\[
\leq 1 - \frac{(1 - 2\mathbb{E}_{v \sim \rho} R_D(G_{Q_v}))^2}{1 - 2\mathbb{E}_{v \sim \rho} d_D(Q_v)},
\]

where \(d_D(\rho)\) is the expected disagreement between pairs of voters defined as

\[
d_D(\rho) = \mathbb{E}_{x \sim D_x} \mathbb{E}_{v \sim \rho} \mathbb{E}_{v' \sim \rho} \mathbb{E}_{h \sim Q_v} \mathbb{E}_{h' \sim Q_{v'}} \mathbb{I}_{[h(x^v) \neq h'(x'^{v'})]}.
\]

and \(R_D(G_{Q_v})\) and \(d_D(Q_v)\) are respectively the true view-specific Gibbs risk and the expected disagreement defined as

\[
R_D(G_{Q_v}) = \mathbb{E}_{(x,y) \sim D} \mathbb{I}_{[h(x^v) \neq y]},
\]

\[
d_D(Q_v) = \mathbb{E}_{x \sim D_x} \mathbb{E}_{h \sim Q_v} \mathbb{I}_{[h(x^v) \neq h'(x'^v)]}.
\]

\(^1\)In the PAC-Bayesian literature, the weighted majority vote is sometimes called the Bayes classifier.
Proof. Similarly than done for the classical C-Bound [Germain et al. 2015, Lacasse et al. 2006], Equation (2) follows from the Cantelli-Chebyshev’s inequality (we provide the proof in [9]). Equation (3) is obtained by rewriting \(R_D(G_\rho)\) as the \(\rho\)-average of the risk associated to each view, and lower-bounding \(d_D(\rho)\) by the \(\rho\)-average of the disagreement associated to each view. First we notice that in the binary setting where \(y \in \{-1, 1\}\) and \(h : \mathcal{X} \to \{-1, 1\}\), we have \(1_{[h(x^v) \neq y]} = \frac{1}{2}(1 - y h(x^v))\), and
\[
R_D(G_\rho) = \mathbb{E}_{(x,y) \sim D} \mathbb{E}_{\nu \sim \rho} \mathbb{E}_{h \sim \nu_{Q_x}} 1_{[h(x^v) \neq y]} = \frac{1}{2} \left( 1 - \mathbb{E}_{(x,y) \sim D} \mathbb{E}_{\nu \sim \rho} y h(x^v) \right) = \mathbb{E}_{\nu \sim \rho} R_D(G_{Q_\nu}).
\]
Moreover, we have
\[
d_D(\rho) = \mathbb{E}_{x \sim D_x} \mathbb{E}_{\nu \sim \rho} \mathbb{E}_{h \sim \nu_{Q_x}} 1_{[h(x^v) \neq h'(x^v)]} = \frac{1}{2} \left( 1 - \mathbb{E}_{x \sim D_x} \mathbb{E}_{\nu \sim \rho} \mathbb{E}_{h \sim \nu_{Q_x}} h(x^v) \times h'(x^v) \right) = \frac{1}{2} \left( 1 - \mathbb{E}_{x \sim D_x} \mathbb{E}_{\nu \sim \rho} h(x^v)^2 \right).
\]
From Jensen’s inequality (Theorem 4 in Appendix) it comes
\[
d_D(\rho) \geq \frac{1}{2} \left( 1 - \mathbb{E}_{x \sim D_x} \mathbb{E}_{\nu \sim \rho} \mathbb{E}_{h \sim \nu_{Q_x}} h(x^v)^2 \right) = \mathbb{E}_{\nu \sim \rho} \left[ \frac{1}{2} \left( 1 - \mathbb{E}_{x \sim D_x} \mathbb{E}_{h \sim \nu_{Q_x}} h(x^v)^2 \right) \right] = \mathbb{E}_{\nu \sim \rho} d_D(Q_\nu).
\]
By replacing \(R_D(G_\rho)\) and \(d_D(\rho)\) in Equation (2), we obtain
\[
1 - \frac{(1 - 2R_D(G_\rho))^2}{1 - 2d_D(\rho)} \leq 1 - \frac{(1 - 2 \mathbb{E}_{\nu \sim \rho} R_D(G_{Q_\nu}))^2}{1 - 2 \mathbb{E}_{\nu \sim \rho} d_D(Q_\nu)}.
\]

Equation (2) suggests that a good trade-off between the Gibbs risk and the disagreement between pairs of voters will lead to a well-performing majority vote. Equation (3) controls the diversity among the views (important for multiview learning [Amini et al. 2009, Goyal et al. 2017, Chapelle et al. 2010, Maillard and Vayatis 2009]) thanks to the disagreement’s expectation over the views \(\mathbb{E}_{\nu \sim \rho} d_D(Q_\nu)\).

3.2 The General Multiview PAC-Bayesian Theorem

In this section, we give a general multiview PAC-Bayesian theorem [Goyal et al. 2017] that takes the form of a generalization bound for the Gibbs risk in the context of a two-level hierarchy of distributions. A key step in PAC-Bayesian proofs is the use of a change of measure inequality [McAllester 2003], based on the Donsker-Varadhan inequality [Donsker and Varadhan 1975]. Lemma 2 below extends this tool to our multiview setting.

Lemma 2 For any set of priors \(\{P_v\}_{v=1}^V\) over \(\{H_v\}_{v=1}^V\) and any set of posteriors \(\{Q_v\}_{v=1}^V\) over \(\{H_v\}_{v=1}^V\), for any hyper-prior distribution \(\pi\) on views \(V\) and hyper-posterior distribution \(\rho\) on \(V\), and for any measurable function \(\phi : \mathcal{H}_v \to \mathbb{R}\), we have
\[
\mathbb{E}_{\nu \sim \pi} \mathbb{E}_{h \sim Q_\nu} \phi(h) \leq \mathbb{E}_{\nu \sim \rho} KL(Q_\nu||P_\nu) + KL(\rho||\pi) + \ln \left( \mathbb{E}_{\nu \sim \pi} \mathbb{E}_{h \sim P_\nu} e^{\phi(h)} \right).
\]

Proof. Deferred to [9].
Theorem 1 Let $V \geq 2$ be the number of views. For any distribution $D$ on $X \times Y$, for any set of prior distributions $\{P_v\}_{v=1}^V$ over $\{\mathcal{H}_v\}_{v=1}^V$, for any hyper-prior distributions $\pi$ over $\mathcal{V}$, for any convex function $D : [0, 1] \times [0, 1] \rightarrow \mathbb{R}$, for any $\delta \in (0, 1]$, with a probability at least $1 - \delta$ over the random choice of $S \sim (D)^n$, for all posterior $\{Q_v\}_{v=1}^V$ over $\{\mathcal{H}_v\}_{v=1}^V$ and hyper-posterior $\rho$ over $D$ distributions, we have:

$$D(R_S(G_\rho), R_D(G_\rho)) \leq \frac{1}{m} \left[ \mathbb{E}_{v \sim \rho} \mathbb{KL}(Q_v \| P_v) \right. + \mathbb{KL}(\rho \| \pi) + \left. \mathbb{KL}(\rho \| \pi) \right] .$$

Proof. First, note that $\mathbb{E}_{v \sim \rho} \mathbb{KL}(Q_v \| P_v)$ is a non-negative random variable. Using Markov’s inequality, with $\delta \in (0, 1]$, and a probability at least $1 - \delta$ over the random choice of the multiview learning sample $S \sim (D)^n$, we have

$$\mathbb{E}_{v \sim \rho} \mathbb{KL}(Q_v \| P_v) \leq \frac{1}{\delta} \mathbb{KL}(\rho \| \pi) .$$

By taking the logarithm on both sides, with a probability at least $1 - \delta$ over $S \sim (D)^n$, we have

$$\mathbb{KL}(\rho \| \pi) \leq \mathbb{KL}(\rho \| \pi) .$$

We now apply Lemma 2 \footnote{Cattani, 2007,} on the left-hand side of Inequality 1 \footnote{Cattani, 2007,} with $\phi(h) = n D(R_S(h), R_D(h))$. Therefore, for any $Q_v$ on $\mathcal{H}_v$ for all views $v \in V$, and for any $\rho$ on views $V$, with a probability at least $1 - \delta$ over $S \sim (D)^n$, we have

$$\mathbb{KL}(\rho \| \pi) \leq \mathbb{KL}(\rho \| \pi) .$$

where the last inequality is obtained by applying Jensen’s inequality on the convex function $D$. By rearranging the terms, we have

$$D(R_S(h), R_D(h)) \leq \frac{1}{m} \left[ \mathbb{E}_{v \sim \rho} \mathbb{KL}(Q_v \| P_v) + \mathbb{KL}(\rho \| \pi) \right] .$$

Finally, the theorem statement is obtained by rewriting

$$\mathbb{E}_{v \sim \rho} R_S(h) = R_S(G_\rho), \quad \mathbb{E}_{v \sim \rho} R_D(h) = R_D(G_\rho).$$

Compared to the classical single-view PAC-Bayesian Bound of \footnote{Goyal, Morvant, Germain, Amini, 2016}, the main difference relies on the introduction of the view-specific prior and posterior distributions, which mainly leads to an additional term $\mathbb{E}_{v \sim \rho} \mathbb{KL}(Q_v \| P_v)$ expressed as the expectation of the view-specific Kullback-Leibler divergence terms over the views $V$ according to the hyper-posterior distribution $\rho$.

Theorem 1 provides tools to derive PAC-Bayesian generalization bounds for a multiview supervised learning setting. Indeed, by making use of the same trick as \footnote{Goyal, Morvant, Germain, Amini, 2016}, by choosing a suitable convex function $D$ and upper-bounding $\mathbb{KL}(\rho \| \pi)$, we obtain an instantiation of Theorem 1. In the next section, we give an example of this kind of deviation through the approach of \footnote{Cattani, 2007}, that is one of the three classical PAC-Bayesian Theorems \footnote{McAllester, 1999,}, \footnote{Cattani, 2007}, \footnote{Seeger, 2002, Langford, 2005}. 

Technical Report V 2 6
3.3 An Example of Instantiation of the Multiview PAC-Bayesian Theorem

To obtain the following theorem which is a generalization bound with the Catoni [2007]’s point of view, we put $D(a, b) = F(b) - C a$ where $F$ is a convex function $F$ and $C > 0$ is a real number [Germain et al. 2009, 2015].

**Corollary 1** Let $V \geq 2$ be the number of views. For any distribution $\mathcal{D}$ on $X \times Y$, for any set of prior distributions $\{P_v\}_{v=1}^V$ on $\mathcal{H}_v$ for any hyper-prior distributions $\pi$ over $\mathcal{V}$, for any $\delta \in (0, 1]$, with a probability at least $1 - \delta$ over the random choice of $S \sim (\mathcal{D})^n$ for all posterior $\{Q_v\}_{v=1}^V$ and hyper-posterior $\rho$ distributions, we have:

$$R_\mathcal{D}(G_\rho) \leq \frac{1}{1 - e^{-\delta}} \left(1 - \exp\left[-\left(C R_S(G_\rho) + \frac{1}{n} \sum_{v=1}^V \text{KL}(Q_v \| P_v) + \text{KL}(\rho \| \pi) + \ln \frac{1}{\delta}\right)\right]\right).$$

**Proof.** Deferred to D.

This bound has the advantage of expressing a trade-off between the empirical Gibbs risk and the Kullback-Leibler divergences.

3.4 A Generalization Bound for the C-Bound

From a practical standpoint, as pointed out before, controlling the multiview C-Bound of Equation (3) can be very useful for tackling multiview learning. The next theorem is a generalization bound that justify the empirical minimization of the multiview C-bound (we use in our algorithm PB-MVBoost derived in Section 4).

**Theorem 2** Let $V \geq 2$ be the number of views. For any distribution $\mathcal{D}$ on $X \times Y$, for any set of prior distributions $\{P_v\}_{v=1}^V$, for any hyper-prior distributions $\pi$ over views $\mathcal{V}$, and for any convex function $D : [0, 1] \times [0, 1] \rightarrow \mathbb{R}$, with a probability at least $1 - \delta$ over the random choice of $S \sim (\mathcal{D})^n$ for all posterior $\{Q_v\}_{v=1}^V$ and hyper-posterior $\rho$ distributions, we have:

$$C \mathcal{D}(B_\rho) \leq 1 - \frac{\left(1 - 2 \frac{\mathbb{E} \sup_{v=1}^V \left(r^{\delta/2}_{Q_v, S}\right)}{1 - 2 \frac{\mathbb{E} \inf_{v=1}^V d^{\delta/2}_{Q_v, S}}{}}\right)^2}{1 - 2 \frac{\mathbb{E} \inf_{v=1}^V \left(d^{\delta/2}_{Q_v, S}\right)}{1 - 2 \frac{\mathbb{E} \sup_{v=1}^V \left(r^{\delta/2}_{Q_v, S}\right)}{}}}.$$

where

$$r^{\delta/2}_{Q_v, S} = \left\{ r : \text{kl}(R_S(Q_v) \| r) \leq \frac{1}{n} \left[ \text{KL}(Q_v \| P_v) + \ln \frac{4\sqrt{m}}{\delta} \right] \text{ and } r \leq \frac{1}{2} \right\}$$

and

$$d^{\delta/2}_{Q_v, S} = \left\{ d : \text{kl}(d^{S}_{Q_v} \| d) \leq \frac{1}{n} \left[ 2 \text{KL}(Q_v \| P_v) + \ln \frac{4\sqrt{m}}{\delta} \right] \right\}.$$

**Proof.** Similarly to Equations (23) and (24) of Germain et al. 2015, we define the sets $r^{\delta/2}_{Q_v, S}$ (Equation (7)) and $d^{\delta/2}_{Q_v, S}$ (Equation (8)) for our setting. Finally, the bound is obtained (from Equation 3 of Lemma 1) by replacing the view-specific Gibbs risk $R_\mathcal{D}(G_\rho)$ by its upper bound $\sup_{v=1}^V \left(r^{\delta/2}_{Q_v, S}\right)$ and expected disagreement $d_\mathcal{D}(Q_v)$ by its lower bound $\inf_{v=1}^V \left(d^{\delta/2}_{Q_v, S}\right)$.

4 The PB-MVBoost algorithm

In this section we exploit our two-level hierarchical strategy (see Figure 4) in order to learn a well-performing weighted combination of view-specific voters (or views) as in Equation (11). Therefore, we propose to follow a well-known approach to learn weighted combination of voters, that is boosting. Indeed, boosting aims at combining a set of weak voters to construct a good majority vote. Typically, boosting algorithms repeatedly learn a “weak” voter (using a learning algorithm) with different probability distribution over the learning sample $S$. Finally, it combines all the weak voters in order to have one single strong classifier.

---

4 In boosting, the performance of a weak classifier is only slightly better than random guessing.
Algorithm 1 PB-MVBoost

Input: Training set \( S = (x_i, y_i), \ldots, (x_n, y_n) \), where \( x_i = (x_i^1, x_i^2, \ldots, x_i^V) \) and \( y_i \in \{-1, 1\} \).

For each view \( v \in V \), a view-specific hypothesis set \( H_v \).

Number of iterations \( T \).

1: for \( x_i \in S \) do
2: \( D_1(x_i) \leftarrow \frac{1}{n} \)
3: \( \forall v \in V \) \( \rho^1_v \leftarrow \frac{1}{V} \)

4: for \( t = 1, \ldots, T \) do
5: \( \forall v \in V \) \( h^t_v \leftarrow \arg \min_{h \in H_v} \mathbb{E}_{(x_i, y_i) \sim D_t} \left[ I[h(x^t_i) \neq y_i] \right] \)
6: Compute error: \( \forall v \in V \) \( \epsilon^t_v \leftarrow \mathbb{E}_{(x_i, y_i) \sim D_t} \left[ I[h^t_v(x^t_i) \neq y_i] \right] \)
7: Compute voter weights (taking into account view specific information): \( \forall v \in V \) \( Q^t_v \leftarrow \frac{1}{2} \left[ \ln \left( \frac{1 - \epsilon^t_v}{\epsilon^t_v} \right) \right] \)

8: Optimize the multiview C-Bound to learn weights over the views

\[
\rho^t \leftarrow \arg \max_{\rho} \left[ \frac{1 - 2 \sum_{v=1}^{V} \rho_v \cdot r^t_v}{1 - 2 \sum_{v=1}^{V} \rho_v \cdot d^t_v} \right]^2 \\
\text{such that} \quad \sum_{v=1}^{V} \rho_v = 1, \quad \rho_v \geq 0 \quad \forall v \in \{1, \ldots, V\} \\
\text{where} \forall v \in V \text{, } r^t_v \leftarrow \mathbb{E}_{(x_i, y_i) \sim D_t} \left[ I[h(x^t_i) \neq y_i] \right] \\
\forall v \in V \text{, } d^t_v \leftarrow \mathbb{E}_{(x_i, y_i) \sim D_t} \left[ I[h(x^t_i) \neq h^t(x^t_i)] \right] \]

9: for \( x_i \in S \) do
10: \( D_{t+1}(x_i) \leftarrow \frac{D_t(x_i) \exp \left( -\gamma_i \sum_{v=1}^{V} \rho^t_v \cdot Q^t_v h^t_v(x^t_i) \right)}{\sum_{j=1}^{n} D_t(x_j) \exp \left( -\gamma_j \sum_{v=1}^{V} \rho^t_v \cdot Q^t_v h^t_v(x^t_j) \right)} \)

11: Return: For each view \( v \in V \), weights over view-specific voters and weights over views, i.e., \( \rho^T \)
which performs better than the individual weak voters. Recall that in multiview learning it is crucial to take into account the interactions between voters and views [Amini et al. 2009, Goyal et al. 2017, Chapelle et al. 2010, Maillard and Vayatis 2009]. We adapt this principle to our setting for combining a set of view-specific weak voters while taking into account the accuracy and diversity between them. We develop a multiview learning algorithm PB-MVBoost (see Algorithm 1), which allows to iteratively learn the set of view-specific classifiers that the algorithm will combine.

Concretely, for a given training set \( S = \{ (x_1, y_1), \ldots, (x_n, y_n) \} \in (\mathcal{X} \times \{ -1, +1 \})^n \) of size \( n \), our algorithm PB-MVBoost maintains a distribution over the examples which is initialized as uniform. Then at each iteration, \( V \) view-specific weak classifiers are learned according to the current distribution \( D_t \) (Step 5), and their corresponding errors \( e^v_t \) are estimated (Step 6).

Similarly to the Adaboost algorithm [Freund and Schapire 1997], the weights of each view-specific classifier \( (Q^v_t)_{1 \leq v \leq V} \) are then computed with respect to these errors as

\[
\forall v \in V, Q^v_t \leftarrow \frac{1}{2} \left[ \ln \left( \frac{1 - e^v_t}{e^v_t} \right) \right].
\]

To learn the weights \( (\rho_v)_{1 \leq v \leq V} \) over the views, we optimize the multiview C-Bound, given by Equation (3) of Lemma 1 (Step 8 of algorithm), which in our case writes as a constraint maximization problem:

\[
\max_{\rho} \quad \frac{1 - 2 \sum_{v=1}^{V} \rho_v r^v_t}{1 - 2 \sum_{v=1}^{V} \rho_v d^v_t},
\]

\[
s.t. \quad \sum_{v=1}^{V} \rho_v = 1, \quad \rho_v \geq 0 \quad \forall v \in \{1, \ldots, V\}.
\]

where \( r^v_t \) is the view-specific Gibbs risk, and \( d^v_t \) the expected disagreement over all view-specific voters defined as follows.

\[
r^v_t = \mathbb{E}_{(x_i, y_i) \sim D_t} \mathbb{E}_{h \sim H_v} \mathbbm{1}[h(x_i^t) \neq y_i],
\]

\[
d^v_t = \mathbb{E}_{(x_i, y_i) \sim D_t} \mathbb{E}_{h, h' \sim H_v} \mathbbm{1}[h(x_i^t) \neq h'(x_i^t)].
\]

Intuitively, the multiview C-Bound tries to diversify the view-specific voters and views (Equation (10)) while controlling the classification error of the view-specific classifiers (Equation (9)). This allows us to control the accuracy and the diversity between the views which is an important ingredient in multiview learning [Xu and Sun 2010, Goyal et al. 2017, Peng et al. 2011, 2017, Morvant et al. 2014].

In Section 5 we empirically show that our algorithm minimizes the multiview C-Bound over the iterations of the algorithm (this is theoretically justified by the generalization bound of Theorem 2). Finally, we update the distribution over training examples \( x_i \) (Step 9), by following the AdaBoost algorithm and in a way that the weights of misclassified (resp. well classified) examples by the final weighted majority classifier increase (resp. decrease).

\[
D_{t+1}(x_i) \leftarrow \frac{D_t(x_i) \exp \left(-y_i \sum_{v=1}^{V} \rho_v Q^v_t h^v_t(x_i^t) \right)}{\sum_{j=1}^{n} D_t(x_j) \exp \left(-y_j \sum_{v=1}^{V} \rho_v Q^v_t h^v_t(x_j^t) \right)}.
\]

Intuitively, this forces the view-specific classifiers to be consistent with each other, which is important for multiview learning [Janodet et al. 2009, Kojo and Capponli 2011, Xiao and Guo 2012]. Finally, after \( T \) iterations of the algorithm, we learn the weights over the view-specific voters and weights over the views leading to a well-performing weighted multiview majority vote defined as

\[
B_\rho(x) = \text{sign} \left( \sum_{v=1}^{V} \rho_v^T \sum_{i=1}^{T} Q^v_i h^v_i(x^v) \right).
\]

4.1 A note on the Complexity of PB-MVBoost

The complexity of learning a decision tree classifier is \( O(d n \log(n)) \), where \( d \) is the depth of the decision tree. We learn the weights over the views by optimizing Equation (3) (Step 8 of our algorithm) using SLSQP method which has time complexity of \( O(V^3) \). Therefore, the overall complexity is \( O \left( T \left( V^3 + V d_n \log(n) \right) \right) \). Note that it is easy to parallelize our algorithm: by using \( V \) different machines, we can learn the view-specific classifiers and weights over them (Steps 4 to 7).
one-vs-all classification problems. While the datasets are multiclass, we transformed them as binary tasks by considering

We reserve 30% of documents as test samples and remaining as training data. The datasets.

Table 1: Test classification accuracy and \( F_1 \)-score of different approaches averaged over all the classes and over 20 random sets of \( n = 500 \) labeled examples per training set. Along each column, the best result is in bold, and second one in italic. \( \dagger \) indicates that a result is statistically significantly worse than the best result, according to a Wilcoxon rank sum test with \( p < 0.02 \).

<table>
<thead>
<tr>
<th>Strategy</th>
<th>MNIST(_1)</th>
<th>MNIST(_2)</th>
<th>Reuters</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Accuracy ( F_1 )</td>
<td>Accuracy ( F_1 )</td>
<td>Accuracy ( F_1 )</td>
</tr>
<tr>
<td>Mono</td>
<td>(-0.9034 \pm 0.01) (\dagger)</td>
<td>(-0.9164 \pm 0.01) (\dagger)</td>
<td>(-0.8420 \pm 0.02) (\dagger)</td>
</tr>
<tr>
<td>Concat</td>
<td>(-0.9224 \pm 0.02) (\dagger)</td>
<td>(-0.9214 \pm 0.02) (\dagger)</td>
<td>(-0.8431 \pm 0.04) (\dagger)</td>
</tr>
<tr>
<td>Fusion(_R)</td>
<td>(-0.9320 \pm 0.01) (\dagger)</td>
<td>(-0.9366 \pm 0.01) (\dagger)</td>
<td>(-0.8587 \pm 0.03) (\dagger)</td>
</tr>
<tr>
<td>MV-MV</td>
<td>(-0.9402 \pm 0.01) (\dagger)</td>
<td>(-0.9450 \pm 0.01) (\dagger)</td>
<td>(-0.8780 \pm 0.02) (\dagger)</td>
</tr>
<tr>
<td>rBoost.SH</td>
<td>(-0.9256 \pm 0.01) (\dagger)</td>
<td>(-0.9545 \pm 0.00) (\dagger)</td>
<td>(-0.8853 \pm 0.02) (\dagger)</td>
</tr>
<tr>
<td>MV-AdaBoost</td>
<td>(-0.9514 \pm 0.01) (\dagger)</td>
<td>(-0.9611 \pm 0.00) (\dagger)</td>
<td>(-0.8942 \pm 0.06) (\dagger)</td>
</tr>
<tr>
<td>MVBoost</td>
<td>(-0.9494 \pm 0.03) (\dagger)</td>
<td>(-0.9555 \pm 0.02) (\dagger)</td>
<td>(-0.8627 \pm 0.07) (\dagger)</td>
</tr>
<tr>
<td>Fusion(_S)</td>
<td>(-0.9418 \pm 0.02) (\dagger)</td>
<td>(-0.9548 \pm 0.03) (\dagger)</td>
<td>(-0.9001 \pm 0.03) (\dagger)</td>
</tr>
<tr>
<td>PB-MVBoost</td>
<td>(-0.9661 \pm 0.00) (\dagger)</td>
<td>(-0.9674 \pm 0.00) (\dagger)</td>
<td>(-0.8953 \pm 0.02) (\dagger)</td>
</tr>
</tbody>
</table>

5 Experimental Results

In this section, we present experiments to show the potential of our algorithm PB-MVBoost on the following datasets.

5.1 Datasets

MNIST

MNIST is a publicly available dataset consisting of 70,000 images of handwritten digits distributed over ten classes [Lecun et al. 1998]. For our experiments, we generated 2 four-view datasets \(\dagger\) where each view is a vector of \(R^{14} \). Similarly than done by Chen and Denoyer [2017], the first dataset (MNIST\(_1\)) is generated by considering 4 quarters of image as 4 views. For the second dataset (MNIST\(_2\)), we consider 4 overlapping views around the centre of images: this dataset brings redundancy between the views. These two datasets allow us to check if our algorithm is able to capture redundancy between the views. We reserve 10,000 of images as test samples and remaining as training samples.

Multilingual, Multiview Text categorization

This dataset is a multilingual text classification data extracted from Reuters RCV1/RCV2 corpus \(\dagger\). It consists of more than 110,000 documents written in five different languages (English, French, German, Italian and Spanish) distributed over six classes. We see different languages as different views of the data. We reserve 30% of documents as test samples and remaining as training data.

5.2 Experimental Protocol

While the datasets are multiclass, we transformed them as binary tasks by considering one-vs-all classification problems: for each class we learn a binary classifier by considering all the learning samples from that class as positive examples and the others as negative examples. We consider different sizes of learning sample \(S\) (150, 200, 300, 500, 800, 1000) that are chosen randomly from the training data. Moreover, all the results are averaged over 20 random runs of the experiments. Since the classes are unbalanced, we report the accuracy along with F1-measure for the methods and all the scores are averaged over all the one-vs-all classification problems.

We consider two multiview learning algorithms based on our two-step hierarchical strategy, and compare the PB-MVBoost \(\dagger\) algorithm described in Section 4 with a previously developed multiview learning algorithm [Goyal et al. 2017], based on classifier late fusion approach [Snoek et al. 2005], and referred to as Fusion\(_S\). Concretely, at the first level, this algorithm trains different view-specific linear SVM models with different hyperparameter \(C\) values (12 values between \(10^{-8}\) and \(10^{3}\)). And, at the second level, it learns a weighted combination over the predictions of view-specific voters using PAC-Bayesian

---

\(\dagger\)MNIST\(_1\) and MNIST\(_2\) datasets are available at [https://github.com/goyalanil/Multiview_Dataset_MNIST](https://github.com/goyalanil/Multiview_Dataset_MNIST)

\(\dagger\)Reuters RCV1/RCV2 corpus is available at [https://archive.ics.uci.edu/ml/datasets/Reuters+RCV1+RCV2](https://archive.ics.uci.edu/ml/datasets/Reuters+RCV1+RCV2)

\(\dagger\)Multilingual+Multiview+Text+Categorization+Test+collection

\(\dagger\)Code for PB-MVBoost is available at [https://github.com/goyalanil/PB-MVBoost](https://github.com/goyalanil/PB-MVBoost)
algorithm CqBoost\cite{Roy2016} with a RBF kernel. Note that, algorithm CqBoost tends to minimize the PAC-Bayesian C-Bound\cite{Germain2015} controlling the trade-off between accuracy and disagreement among voters. The hyperparameter $\gamma$ of the RBF kernel is chosen over a set of 9 values between $10^{-6}$ and $10^{0}$; and hyperparameter $\mu$ is chosen over a set of 8 values between $10^{-8}$ and $10^{-1}$. To study the potential of our algorithms (\text{Fusion}^{\text{all}}_{\text{Cq}} and \text{PB-MVBoost}), we considered following 7 baseline approaches:

- **Mono**: We learn a view-specific model for each view using a decision tree classifier and report the results of the best performing view.

- **Concat**: We learn one model using a decision tree classifier by concatenating features of all the views.

- **\text{Fusion}^{\text{all}}_{\text{Cq}}**: This is a late fusion approach where we first learn the view-specific classifiers using 60\% of learning samples. Then, we learn a final multiview weighted model over the predictions of the view-specific classifiers. For this approach, we used decision tree classifiers at both levels of learning.

- **MV-MV**: We compute a multiview uniform majority vote (similar to approach followed by \text{Amini} et al. \cite{Amini2009}) over all the view-specific classifiers’ outputs in order to make final prediction. We learn view-specific classifiers using decision tree classifiers.

- **rBoost.SH**: This is the multiview learning algorithm proposed by \text{Peng} et al. \cite{Peng2011,Peng2017} where a single global distribution is maintained over the learning sample for all the views and the distribution over views are updated using multiarmed bandit framework. At each iteration, rBoost.SH selects a view according to the current distribution and learns the corresponding view-specific voter. For tuning the parameters, we followed the same experimental setting as \text{Peng} et al. \cite{Peng2017}.

- **MV-AdaBoost**: This is a majority vote classifier over the view-specific voters trained using Adaboost algorithm. Here, our objective is to see the effect of maintaining separate distributions for all the views.

- **MVBoost**: This is a variant of our algorithm PB-MVBoost but without learning weights over views by optimizing multiview C-Bound. Here, our objective is to see the effect of learning weights over views on multiview learning.

For all boosting based approaches (rBoost.SH, MV-AdaBoost, MVBoost and PB-MVBoost), we learn the view-specific voters using a decision tree classifier with depth 2 and 4 as a weak classifier for MNIST, and Reuters RCV1/RCV2 datasets respectively. For all these approaches, we kept $T = 100$ as the number of iterations. For optimization of multiview C-Bound, we used Sequential Least Squares Programming (SLSQP) implementation provided by SciPy\cite{Jones2001} and the decision trees implementation from scikit-learn\cite{Pedregosa2011}.

### 5.3 Results

Firstly, we report the comparison of our algorithms Fusion$^{\text{all}}_{\text{Cq}}$ and PB-MVBoost (for $m = 500$) with all the considered baseline methods in Table 1. Secondly, Figure 2 illustrates the evolution of the performances according to the size of the learning sample. From the table, proposed two-step learning algorithm Fusion$^{\text{all}}_{\text{Cq}}$ is significantly better than the baseline approaches for Reuters dataset. Whereas, our boosting based algorithm PB-MVBoost is significantly better than all the baseline approaches for all the datasets. This shows that considering a two-level hierarchical strategy in a PAC-Bayesian manner is an effective way to handle multiview learning.

In Figure 3, we compare proposed algorithms Fusion$^{\text{all}}_{\text{Cq}}$ and PB-MVBoost in terms of accuracy, $F_1$-score and time complexity for $m = 500$ examples. For MNIST datasets, PB-MVBoost is significantly better than Fusion$^{\text{all}}_{\text{Cq}}$. For Reuters dataset, Fusion$^{\text{all}}_{\text{Cq}}$ performs better than PB-MVBoost but computation time for Fusion$^{\text{all}}_{\text{Cq}}$ is much higher than that of PB-MVBoost. Moreover, in Figure 2 we can see that the performance (in terms of $F_1$-score) for Fusion$^{\text{all}}_{\text{Cq}}$ is worse than PB-MVBoost when we have less training examples ($n = 150$ and 200). This shows the proposed boosting based one-step algorithm PB-MVBoost is more stable and more effective for multiview learning.

From Table 1 and Figure 2, we can observe that MV-AdaBoost (where we have different distributions for each view over the learning sample) provides better results compared to other baselines in terms of accuracy but not in terms of $F_1$-measure. On the other hand, MVBoost (where we have single global
distribution over the learning sample but without learning weights over views) is better compared to other baselines in terms of F1-measure. Moreover, the performances of MVBoost first increases with an increase of the quantity of the training examples, then decreases. Whereas our algorithm PB-MVBoost provides the best results in terms of both accuracy and F1-measure, and leads to a monotonic increase of the performances with respect to the addition of labeled examples. This confirms that by maintaining a single global distribution over the views and learning the weights over the views using a PAC-Bayesian framework, we are able to take advantage of different representations (or views) of the data.

Finally, we plot behaviour of our algorithm PB-MVBoost over $T = 100$ iterations on Figure 4 for all the datasets. We plot accuracy and F1-measure of learned models on training and test data along with empirical multiview C-Bound on training data at each iteration of our algorithm. Over the iterations, the F1-measure on the test data keeps on increasing for all the datasets even if F1-measure and accuracy on the training data reach the maximal value. This confirms that our algorithm handles unbalanced data well. Moreover, the empirical multiview C-Bound (which controls the trade-off between accuracy and diversity between views) keeps on decreasing over the iterations. This validates that by combining the PAC-Bayesian framework with the boosting one, we can empirically ensure the view specific information and diversity between the views for multiview learning.
6 Conclusion

In this paper, we provide a PAC-Bayesian analysis for a two-level hierarchical multiview learning approach with more than two views, when the model takes the form of a weighted majority vote over a set of functions/voters. We consider a hierarchy of weights modelled by distributions where for each view we aim at learning i) posterior $Q_v$ distributions over the view-specific voters capturing the view-specific information and ii) hyper-posterior $\rho_v$ distributions over the set of the views. Based on this strategy, we derived a general multiview PAC-Bayesian theorem that can be specialized to any convex function to compare the empirical and true risks of the stochastic multiview Gibbs classifier. We propose a boosting-based learning algorithm, called as PB-MVBoost. At each iteration of the algorithm, we learn the weights over the view-specific voters and the weights over the views by optimizing an upper-bound over the risk of the majority vote (the multiview C-Bound) that has the advantage of controlling a trade-off between accuracy and the diversity between the views. The empirical evaluation shows that PB-MVBoost leads to good performances and confirms that our two-level PAC-Bayesian strategy is indeed a nice way to tackle multiview learning. Moreover, we compare the effect of maintaining separate distributions over the learning sample for each view; single global distribution over views; and single global distribution along with learning weights over views on results of multiview learning. We show that by maintaining a single global distribution over the learning sample for all the views and learning the weights over the views is an effective way to deal with multiview learning. In this way, we are able to capture the view-specific information and control the diversity between the views. Finally, we compare PB-MVBoost with a two-step learning algorithm Fusion$^{all}_q$ which is based on PAC-Bayesian theory. We show that PB-MVBoost is more stable and computationally faster than Fusion$^{all}_q$.

For future work, we would like to specialize our PAC-Bayesian generalization bounds to linear classifiers [Germain et al. 2009] which will clearly open the door to derive theoretically founded multiview learning algorithms. We would also like to extend our algorithm to semi-supervised multiview learning where one has access to an additional unlabeled data during training. One possible way is to learn a view-specific voter using pseudo-labels (for unlabeled data) generated from the voters trained from other views (as done for example in [Xu et al. 2016]). Another possible direction is to make use of unlabeled data while computing view-specific disagreement for optimizing multiview C-Bound. This clearly opens the door to derive theoretically founded algorithms for semi-supervised multiview learning using PAC-Bayesian theory. We would like to extend our algorithm to transfer learning setting where training and test data are drawn from different distributions. An interesting direction would be to bind the data distribution to the different views of the data, as in some recent zero-shot learning approaches [Socher et al. 2013]. Moreover, we would like to extend our work to the case of missing views or incomplete views e.g. Amini et al. 2009 and Xu et al. 2015. One possible solution is to learn the view-specific voters using available view-specific training examples and adapt the distribution over the learning sample accordingly.
Figure 4: Plots for classification error and F1-measure on training and test data; and empirical multiview C-Bound on training data over the iterations for all datasets with \( n = 500 \).
Acknowledgements

This work was partially funded by the French ANR project LIVES ANR-15-CE23-0026-03 and the “Région Rhône-Alpes”.

Appendix

A Mathematical Tools

Theorem 3 (Markov’s ineq.) For any random variable $X$ s.t. $E(|X|) = \mu$, for any $a > 0$, we have $P(|X| \geq a) \leq \frac{\mu}{a}$.

Theorem 4 (Jensen’s ineq.) For any random variable $X$, for any concave function $g$, we have $g(E[X]) \geq E[g(X)]$.

Theorem 5 (Cantelli-Chebyshev ineq.) For any random variable $X$ s.t. $E(X) = \mu$ and $\text{Var}(X) = \sigma^2$, and for any $a > 0$, we have $P(X - \mu \geq a) \leq \frac{\sigma^2}{\mu^2 + \sigma^2}$.

B Proof of $\mathcal{C}$-Bound for Multiview Learning (Lemma 1)

In this section, we present the proof of Lemma 1 inspired by the proof provided by Germain et al. [2015]. Firstly, we need to define the margin of the multiview weighted majority vote $B_\rho$ and its first and second statistical moments.

Definition 1 Let $M_\rho$ is a random variable that outputs the margin of the multiview weighted majority vote on the example $(x, y)$ drawn from distribution $\mathcal{D}$, given by:

$$M_\rho(x, y) = \mathbb{E}_{v \sim \rho} \mathbb{E}_{h \sim Q_v} y h(x^v).$$

The first and second statistical moments of the margin are respectively given by

$$\mu_1(M_\rho) = \mathbb{E}_{(x, y) \sim \mathcal{D}} M_\rho(x, y).$$

and,

$$\mu_2(M_\rho) = \mathbb{E}_{(x, y) \sim \mathcal{D}} [M_\rho(x, y)]^2$$

$$= \mathbb{E}_{x \sim \mathcal{D}_x} y^2 \left[ \mathbb{E}_{v \sim \rho} \mathbb{E}_{h \sim Q_v} h(x^v) \right]^2 = \mathbb{E}_{x \sim \mathcal{D}_x} \left[ \mathbb{E}_{v \sim \rho} \mathbb{E}_{h \sim Q_v} h(x^v) \right]^2. \quad (12)$$

According to this definition, the risk of the multiview weighted majority vote can be rewritten as follows:

$$R_\mathcal{D}(B_\rho) = \mathbb{P}_{(x, y) \sim \mathcal{D}} (M_\rho(x, y) \leq 0).$$

Moreover, the risk of the multiview Gibbs classifier can be expressed thanks to the first statistical moment of the margin. Note that in the binary setting where $y \in \{-1, 1\}$ and $h : \mathcal{X} \rightarrow \{-1, 1\}$, we have $\mathbb{I}_{[h(x^v) \neq y]} = \frac{1}{2} (1 - y h(x^v))$, and therefore

$$R_\mathcal{D}(G_\rho) = \mathbb{E}_{(x, y) \sim \mathcal{D}} \mathbb{E}_{v \sim \rho} \mathbb{E}_{h \sim Q_v} \mathbb{I}_{[h(x^v) \neq y]}$$

$$= \frac{1}{2} \left( 1 - \mathbb{E}_{(x, y) \sim \mathcal{D}} \mathbb{E}_{v \sim \rho} \mathbb{E}_{h \sim Q_v} y h(x^v) \right)$$

$$= \frac{1}{2} (1 - \mu_1(M_\rho)). \quad (13)$$

Similarly, the expected disagreement can be expressed thanks to the second statistical moment of the margin by

$$d_\mathcal{D}(\rho) = \mathbb{E}_{x \sim \mathcal{D}_x} \mathbb{E}_{v \sim \rho} \mathbb{E}_{v' \sim \rho} \mathbb{E}_{h \sim Q_v} \mathbb{E}_{h' \sim Q_{v'}} \mathbb{I}_{[h(x^v) \neq h'(x^{v'})]}$$
From above, we can easily deduce that \(0 \leq d_D(\rho) \leq 1/2\) as \(0 \leq \mu_2(M_P^D) \leq 1\). Therefore, the variance of the margin can be written as:

\[
\Var(M_P^D) = \Var_{(x,y) \sim D}(M_P(x,y)) = \mu_2(M_P^D) - (\mu_1(M_P^D))^2.
\] (15)

The proof of the \(C\)-bound

**Proof.** By making use of one-sided Chebyshev inequality (Theorem 5 of [X]), with \(X = -M_P(x,y)\), \(\mu = \E_{(x,y) \sim D}(M_P(x,y))\) and \(a = \E_{(x,y) \sim D} M_P(x,y)\), we have

\[
R_D(B_\rho) = \Prob_{(x,y) \sim D}(M_P(x,y) \leq 0)
\]

\[
= \Prob_{(x,y) \sim D} \left( -M_P(x,y) + \E_{(x,y) \sim D} M_P(x,y) \geq \E_{(x,y) \sim D} M_P(x,y) \right)
\]

\[
\leq \frac{\Var(M_P(x,y))}{\Var_{(x,y) \sim D}(M_P(x,y)) + \left( \E_{(x,y) \sim D} M_P(x,y) \right)^2}
\]

\[
= \frac{\Var(M_P^D)}{\mu_2(M_P^D) - (\mu_1(M_P^D))^2 + (\mu_1(M_P^D))^2}
\]

\[
= \frac{\Var(M_P^D)}{\mu_2(M_P^D) - (\mu_1(M_P^D))^2}
\]

\[
= \frac{\mu_2(M_P^D) - (\mu_1(M_P^D))^2}{\mu_2(M_P^D)}
\]

\[
= 1 - \frac{\left( \mu_1(M_P^D) \right)^2}{\mu_2(M_P^D)}
\]

\[
= 1 - \frac{\left( \mu_1(M_P^D) \right)^2}{1 - 2d_D(\rho)}
\]

C Proof of Lemma 2

We have

\[
\E_{v \sim p \ h \sim Q_v} \phi(h) = \E_{v \sim p \ h \sim Q_v} \ln(e^{\phi(h)})
\]

\[
= \E_{v \sim p \ h \sim Q_v} \ln \left( \frac{Q_v(h) P_v(h)}{P_v(h) Q_v(h)} e^{\phi(h)} \right)
\]
According to the Kullback-Leibler definition, we have

\[
E_{v \sim \rho} E_{h \sim Q_v} \phi(h) = E_{v \sim \rho} \left[ KL(Q_v \| P_v) + E_{h \sim Q_v} \ln \left( \frac{P_v(h)}{Q_v(h)} e^{\phi(h)} \right) \right].
\]

By applying Jensen’s inequality (Theorem 4 in Appendix) on the concave function \( \ln \), we have

\[
E_{v \sim \rho} E_{h \sim Q_v} \phi(h) \leq E_{v \sim \rho} \left[ KL(Q_v \| P_v) + \ln \left( E_{h \sim P_v} e^{\phi(h)} \right) \right]
\]

\[
= E_{v \sim \rho} KL(Q_v \| P_v) + E_{v \sim \rho} \left( \frac{\rho(v)}{\pi(v)} \pi(v) e^{\phi(h)} \right)
\]

\[
= E_{v \sim \rho} KL(Q_v \| P_v) + KL(\rho \| \pi) + E_{v \sim \rho} \left( \frac{\pi(v)}{\rho(v)} e^{\phi(h)} \right).
\]

Finally, we apply again the Jensen inequality (Theorem 4) on \( \ln \) to obtain the lemma.

D A Catoni-Like Theorem—Proof of Corollary 1

The result comes from Theorem 1 by taking \( D(a,b) = F(b) - Ca \), for a convex \( F \) and \( C > 0 \), and by upper-bounding \( E_{v \sim \rho} E_{h \sim P_v} e^{\rho(D(R_S(h), R_{\pi}(h)))} \). We consider \( R_S(h) \) as a random variable following a binomial distribution of \( n \) trials with a probability of success \( R(h) \). We have:

\[
E_{S \sim (D)^n} E_{\pi \sim \mu} E_{h \sim P_v} e^{n D(R_S(h), R_{\pi}(h))}
\]

\[
= E_{S \sim (D)^n} E_{\pi \sim \mu} E_{h \sim P_v} e^{n F(R_S(h)) - C n R_S(h))}
\]

\[
= E_{S \sim (D)^n} E_{\pi \sim \mu} E_{h \sim P_v} e^{n F(R_S(h))} \sum_{k=0}^{n} P_{R}(R_S(h) = k) e^{-Ck}
\]

\[
= E_{S \sim (D)^n} E_{\pi \sim \mu} E_{h \sim P_v} e^{n F(R_S(h))} \left( R_{\pi}(h) e^{-C} + (1 - R_{\pi}(h)) \right)^n.
\]

The corollary is obtained with \( F(p) = \ln \frac{1}{(1-p)(1-e^{-C})} \).

References


