Rademacher Complexity Bounds for a Penalized Multiclass Semi-Supervised Algorithm (Extended Abstract)*

Yury Maximov[1,2], **Massih-Reza Amini**[3] and **Zaid Harchaoui**[4]
[1] Skolkovo Institute of Science and Technology, Center for Energy Systems
[2] Los Alamos National Laboratory, Theoretical Division T-4 and CNLS
[4] Université Grenoble Alpes
[3] University of Washington
yury@lanl.gov, massih-reza.amini@imag.fr, zaid@uw.edu

Abstract

We propose Rademacher complexity bounds for multi-class classifiers trained with a two-step semi-supervised model. In the first step, the algorithm partitions the partially labeled data and then identifies dense clusters containing κ predominant classes using the labeled training examples such that the proportion of their non-predominant classes is below a fixed threshold stands for clustering consistency. In the second step, a classifier is trained by minimizing a margin empirical loss over the labeled training set and a penalization term measuring the disability of the learner to predict the κ predominant classes of the identified clusters. The resulting data-dependent generalization error bound involves the margin distribution of the classifier, the stability of the clustering technique used in the first step and Rademacher complexity terms corresponding to partially labeled training data. Our theoretical result exhibit convergence rates extending those proposed in the literature for the binary case, and experimental results show empirical evidence that supports the theory.

1 Introduction

Learning with partially labeled data, or Semi-supervised learning (SSL), has been an active field of study in the ML community these past twenty years. In this case, labeled examples are usually supposed to be very few leading to an inefficient supervised model, while unlabeled training examples contain valuable information on the prediction problem at hand which exploitation may lead to a performant prediction function. For this scenario, we assume available a set of labeled training examples $S_\ell = (\mathbf{x}_i, y_i)_{1\leq i\leq n} \in (\mathcal{X}\times\mathcal{Y})^n$ drawn i.i.d. with respect to a fixed, but unknown, probability distribution $\mathcal{D}$ over $\mathcal{X}\times\mathcal{Y}$ and a set of unlabeled training examples $S_u = (\mathbf{x}_{n+i})_{1\leq i\leq u} \in \mathcal{X}^u$ supposed to be drawn from the marginal distribution, $\mathcal{D}_\mathcal{X}$, over the domain $\mathcal{X}$. If S_u is empty, then the problem is cast into the supervised learning framework. The other extreme case corresponds to the situation where S_ℓ is empty and for which the problem reduces to unsupervised learning.

The issue of learnability with partially labeled data was studied under three related yet different hypotheses of *smoothness assumption*, *cluster assumption*, and *low density separation* [Chapelle *et al.*, 2006; Zhu, 2005] and many advances have been made on both algorithmic and theoretical front under these settings.

Although classification problems, for which the design of SSL techniques is appealing, are multi-class in nature, the majority of theoretical results for semi-supervised learning has mainly considered the binary case [Kääriäinen, 2005; Leskes, 2005; Amini *et al.*, 2008; El-Yaniv and Pechyony, 2009; Balcan and Blum, 2010; Urner *et al.*, 2011]. In this paper, we tackle the learning ability of multi-class classifiers trained on partially labeled data by first identifying dense clusters covering labeled and unlabeled examples and then minimizing an objective composed of the margin empirical loss of the classifier over the labeled training set, and also a penalization term measuring the disability of the learner to predict the predominant classes of dense clusters.

Our main result is a data-dependent generalization error bound for classifiers trained under this setting and which exhibits a complexity term depending on the effectiveness of the clustering technique to find homogenous regions of examples belonging to each class, the margin distribution of the classifiers and the Rademacher complexities of the class of functions in use defined for labeled and unlabeled data. The convergence rates deduced from the bound extends those proposed in the literature for the binary case, further experiments carried out on text and image classification problems, show that the proposed approach yields improved classification performance compared to extensions of state-of-the-art SSL algorithms to the multi-class classification case.

In the following section, we first define our framework, then the learning task we address. Section 3 presents the Rademacher generalization bound for a classifier trained with the proposed algorithm. Section 4 positions our theoretical

*This paper is an extended abstract of an article in the Journal of Artificial Intelligence Research [Maximov *et al.*, 2018]

findings corcerning the state-of-the-art, and concludes this work.

2 Framework and Definitions

We are interested in the study of multi-class classification problems where the output space is $\mathcal{Y} = \{1, \ldots, K\}$, with $K > 2$. The semi-supervised multi-class classification algorithm that we consider is tailored under the cluster assumption and operates in two steps depicted in the following sections.

2.1 Partitioning of Data and Identifying κ-Uniformly Bounded Clusters with Level η

The first step consists in partitioning the unlabeled training observations, into $G > 0$ separate clusters with a clustering algorithm $\mathcal{A}$ trained on S_u, denoted by Π_{S_u}.

Clusters of Π_{S_u} that are well covered by classes in the labeled training set are then kept for learning the classifier (Section 2.2). Formally, for a fixed $\kappa \in \{1, \ldots, K\}$, let $\mathcal{Y}_\kappa(\mathcal{C})$ be the κ most predominant classes of $\mathcal{Y}$ present in cluster $\mathcal{C} \in \Pi_{S_u}$. We then define κ-uniformly bounded clusters with level η, $\mathcal{C}_\kappa(\eta)$, the set of clusters within Π_{S_u} that are covered by their κ most predominant classes such that the proportion of other classes within $\mathcal{C}$ not belonging to $\mathcal{Y}_\kappa(\mathcal{C})$ is less than η/G :

$$\mathcal{C}_\kappa(\eta) = \left\{\mathcal{C} \in \Pi_{S_u} \colon P_n((\mathbf{x}, y) \in \mathcal{C} \wedge y \in \mathcal{Y} \backslash \mathcal{Y}_\kappa(\mathcal{C})) \leq \eta/G\right\}.$$

Where P_n the uniform probability distribution over S_ℓ; is defined for any subset $B \subseteq S_\ell$, as $P_n(B) = \text{card}(B)/n$.

2.2 Learning Objective

In the second step, we address a learning problem that is to find, in a hypothesis set $\mathcal{H} \subseteq \mathbb{R}^{\mathcal{X}\times\mathcal{Y}}$, a scoring function $h \in \mathcal{H}$ with low risk:

$$R(h) = \mathbb{E}_{(\mathbf{x},y)\sim\mathcal{D}}\left[1_{m_h(\mathbf{x},y)\leqslant 0}\right],$$

where 1_π is the indicator function and $m_h(\mathbf{x}, y)$ is the margin of the function h at an example $(\mathbf{x}, y)$ [Koltchinskii and Panchenko, 2002]:

$$m_h(\mathbf{x}, y) = h(\mathbf{x}, y) - \max_{y' \in \mathcal{Y}\backslash\{y\}} h(\mathbf{x}, y').$$

This is achieved by minimizing a penalized empirical loss, defined for a given $\rho > 0$:

$$\widehat{R}_\rho(h) = \widehat{R}_\rho(h, S_\ell) + \Omega_\rho(h, \mathcal{C}_\kappa(\eta)), \tag{1}$$

composed of an empirical margin loss of $h \in \mathcal{H}$ on a labeled training set S_ℓ

$$\widehat{R}_\rho(h, S_\ell) = \frac{1}{n} \sum_{(\mathbf{x},y)\in S_\ell} \Phi_\rho(m_h(\mathbf{x}, y)),$$

and a penalization term that reflects the ability of the hypothesis $h \in \mathcal{H}$ to identify the κ most predominant classes within the disjoint clusters of $\mathcal{C}_\kappa(\eta)$:

$$\Omega_\rho(h, \mathcal{C}_\kappa(\eta)) = \frac{1}{u} \sum_{\mathcal{C}\in\mathcal{C}_\kappa(\eta)} \sum_{\mathbf{x}\in\mathcal{C}} \Phi_\rho(m'_h(\mathbf{x}, \mathcal{Y}_\kappa(\mathcal{C}))),$$

where $m'_h(\mathbf{x}, \mathcal{Y}_\kappa(\mathcal{C}))$ is the margin of an unlabeled example taken with respect to the set of κ predominant classes, $\mathcal{Y}_\kappa(\mathcal{C})$:

$$m'_h(\mathbf{x}, \mathcal{Y}_\kappa(\mathcal{C})) = \max_{y\in\mathcal{Y}_\kappa(\mathcal{C})} h(\mathbf{x}, y) - \max_{y\in\mathcal{Y}\backslash\mathcal{Y}_\kappa(\mathcal{C})} h(\mathbf{x}, y),$$

and, $\Phi_\rho : \mathbb{R} \to [0, 1]$ is the ρ-margin loss defined as [Koltchinskii and Panchenko, 2002] :

$$\forall z \in \mathbb{R}, \Phi_\rho(z) = \begin{cases} 0 & \text{if } \rho \geqslant z, \\ 1 - z/\rho & \text{if } 0 < z < \rho, \\ 1 & \text{if } z \leqslant 0. \end{cases}$$

The pseudo-code of the proposed 2-step approach, referred to as Penalized Multi-Class Semi-Supervised Learning ($\texttt{PMS}_2\texttt{L}$), is given in Algorithm 1.

Algorithm 1: Pseudo-code of the $\texttt{PMS}_2\texttt{L}$ algorithm

Input: Labeled data set $S_\ell = (\mathbf{x}_i, y_i)_{1\leq i\leq n} \subseteq (\mathcal{X} \times \mathcal{Y})^n$;
Unlabeled data set $S_u = (\mathbf{x}_{n+i})_{1\leq i\leq u} \subseteq \mathcal{X}^u$;
Hypothesis space $\mathcal{H}$;
G the number of clusters, $\mathcal{A}_{S_u} : \mathcal{X} \to \{1, \ldots, G\}$ the clustering algorithm found on S_u, $\kappa \in \mathbb{N}^*$, and $\eta > 0$;
Stage 1: Using the labeled examples, S_ℓ, identify the κ-bounded clusters in Π_{S_u} with level $\eta, \mathcal{C}_\kappa(\eta)$; // in accordance with the definition of $\mathcal{C}_\kappa(\eta)$
Stage 2: Find a hypothesis $h^* \in \mathcal{H}$ that minimizes the penalized objective function (Eq. 1) :

$$h^* = \operatorname*{argmin}_{h\in\mathcal{H}} \widehat{R}_\rho(h)$$

Output: h^*

3 Theoretical Study

We now analyze how the use of unlabeled training data can improve generalization performance in some cases. Essentially, the trade-off is that clustering offers additional knowledge on the problem, therefore potentially helps to learn, but can also be of lower quality, which may degrade it.

3.1 Stable Clustering with the Bounded Difference Property

Before, let us first introduce notations that are used in the statement of the following results. We consider a hard clustering algorithm $\mathcal{A}_Z$ defined as a function found over a finite sample Z.

Our analyzes are based on a notion of stability of the clustering algorithm $\mathcal{A}_\cdot$; measured as the average number of examples in a given set $\tilde{Z}$ of size n that are in the exclusive disjunction of clusters (present in one and absent from the other) found by $\mathcal{A}_\cdot$ over two sets Z and Z', and defined as :

$$\Delta_n\left(\mathcal{A}_Z, \mathcal{A}_{Z'}, \tilde{Z}\right) = \min_\pi \left[\frac{1}{n} \sum_{\mathbf{x}\in\tilde{Z}} 1_{\mathcal{A}_Z(\mathbf{x})\neq\pi(\mathcal{A}_{Z'}(\mathbf{x}))}\right],$$

where $\pi : \{1, \ldots, G\} \to \{1, \ldots, G\}$ is a permutation. It is straightforward to show that Δ_n defines a true metric, sometimes referred to as the minimal matching distance [Luxburg, 2010], on the space of clusterings. Hence, the clustering algorithm $\mathcal{A}_\cdot$ is said to obey the bounded difference property, if

and only if for any i.i.d. samples Z, $Z' \sim \mathcal{D}_{\mathcal{X}}^{|Z|}$ differing in exactly one observation, and for any i.i.d. sample $\tilde{Z} \sim \mathcal{D}_{\mathcal{X}}^{n}$ of size n, there exists a universal constant L such that :

$$\Delta(\mathcal{A}_Z, \mathcal{A}_{Z'}) = \mathbb{E}_{\tilde{Z}\sim\mathcal{D}_{\mathcal{X}}^{n}}\left[\Delta_n\left(\mathcal{A}_Z, \mathcal{A}_{Z'}, \tilde{Z}\right)\right] \leqslant \frac{L}{|Z|}.$$

For some clustering algorithms such as k-means or k-hyperplane clustering, it has been shown that the bounded difference property is tightly related to their (in)stability. We refer to results of [Luxburg, 2010], [Luxburg *et al.*, 2004], [Rakhlin and Caponnetto, 2006] and [Thiagarajan *et al.*, 2011] and a number of references therein for the algorithmic details as well as various notions of clustering instability, and to that of [Shamir and Tishby, 2007] for the relation between bounded differences property, stability and model selection.

Furthermore, in the case where a clustering algorithm $\mathcal{A}$ obeys the bounded difference property; it is said to be stable if for any distribution $\mathcal{D}_{\mathcal{X}}$ over $\mathcal{X}$ there exists a unique limit clustering of the input space $\Pi^\star$, obtained by a particular instantiation of the algorithm denoted by $\mathcal{A}^\star$, such that for any Z drawn i.i.d. from $\mathcal{D}_{\mathcal{X}}$ and for any sample $\tilde{Z}$ of size n drawn i.i.d. from the same distribution we have :

$$\mathbb{E}_{Z\sim\mathcal{D}_{\mathcal{X}}^{|Z|}}\left[\Delta(\mathcal{A}_Z, \mathcal{A}^\star)\right] \leqslant \frac{L}{|Z|}.$$

In this case, it is possible to (tightly) upper-bound the distance between $\mathcal{A}^\star$ and the algorithm $\mathcal{A}$ trained on any unlabeled training set S_u, estimated over the labeled training set S_ℓ: $\Delta_n(\mathcal{A}_{S_u}, \mathcal{A}^\star, S_\ell)$, as it is stated in the following Lemma.

Lemma 1 *Let $S_\ell = (\mathbf{x}_i, y_i)_{1\leq i\leq n}$ and $S_u = (\mathbf{x}_{n+i})_{1\leq i\leq u}$ be a labeled and an unlabeled training sets drawn i.i.d. according respectively to a probability distribution $\mathcal{D}$ over $\mathcal{X}\times\mathcal{Y}$, and its marginal $\mathcal{D}_{\mathcal{X}}$. For any $1 > \delta > 0$ and any stable clustering algorithm $\mathcal{A}$ that obeys the bounded differences property with constant $L > 0$, the average number of examples in S_ℓ that are in the exclusive disjunction of clusters found by the clustering algorithm $\mathcal{A}$ on S_u and by $\mathcal{A}^\star$ is upper-bounded with probability at least $1-\delta$ as follows :*

$$\Delta_n(\mathcal{A}_{S_u}, \mathcal{A}^\star, S_\ell) \leqslant \frac{L}{u} + L\sqrt{\frac{\log\frac{2}{\delta}}{2u}} + \sqrt{\frac{\log\frac{2}{\delta}}{2n}}.$$

The proof is given in [Maximov *et al.*, 2018]. This result suggests that for any labeled and unlabeled training data, if a clustering algorithm obeys the bounded differences property and that it is stable, then with high probability, Π_{S_u} covers as well the labeled training data as the limit partition $\Pi^\star$ (i.e. most of the labeled examples would more likely be present in the intersection $\Pi_{S_u}\cap\Pi^\star$).

3.2 Semi-Supervised Data-Dependent Bounds

Based on the previous lemma, we can define situations where the Empirical Risk Minimization principle of algorithm PMS$_2$L becomes consistent. This result is stated in Theorem (2) which provides bounds on the generalization error of a multi-class classifier trained with the penalized empirical loss defined above (Eq. 1).

The notion of function class capacity used in the bounds, is the labeled and unlabeled Rademacher complexities of the function class $\mathcal{F}_{\mathcal{H}} = \{f : \mathbf{x} \mapsto h(\mathbf{x}, y) : y \in \mathcal{Y}, h \in \mathcal{H}\}$, defined respectively as:

$$\mathfrak{R}_n^*(\mathcal{F}_{\mathcal{H}}) = \sum_{\mathcal{C}\in\mathcal{C}_\kappa(\eta)} \mathbb{E}_\sigma \sup_{f\in\mathcal{F}_{\mathcal{H}}} \frac{2}{n} \sum_{\mathbf{x}_i\in S_\ell\cap\mathcal{C}} \sigma_i f(\mathbf{x}_i),$$

$$\mathfrak{R}_u^*(\mathcal{F}_{\mathcal{H}}) = \sum_{\mathcal{C}\in\mathcal{C}_\kappa(\eta)} \mathbb{E}_\sigma \sup_{f\in\mathcal{F}_{\mathcal{H}}} \frac{2}{u} \sum_{\mathbf{x}_i\in S_u\cap\mathcal{C}} \sigma_i f(\mathbf{x}_i),$$

$$\mathfrak{R}_n(\mathcal{F}_{\mathcal{H}}) = \mathbb{E}_\sigma \sup_{f\in\mathcal{F}_{\mathcal{H}}} \frac{2}{n} \sum_{\mathbf{x}_i\in S_\ell\setminus\mathcal{C}_\kappa(\eta)} \sigma_i f(\mathbf{x}_i)$$

where σ_i's, called Rademacher variables, are independent uniform random variables taking values in $\{-1, +1\}$; i.e. $\forall i, \mathbb{P}(\sigma_i = -1) = \mathbb{P}(\sigma_i = +1) = \frac{1}{2}$.

The proof is provided in [Maximov *et al.*, 2018]. From this result and Lemma 1, we can then derive a data-dependent generalization bound for any semi-supervised multi-class prediction function found by algorithm PMS$_2$L as stated below.

Theorem 2 *Let $\mathcal{H} \subseteq \mathbb{R}^{\mathcal{X}\times\mathcal{Y}}$ be a hypothesis set where $\mathcal{Y} = \{1, \ldots, K\}$, and let $S_\ell = ((\mathbf{x}_i, y_i))_{i=1}^n$ and $S_u = (\mathbf{x}_i)_{i=n+1}^{n+u}$ be two sets of labeled and unlabeled training data, drawn i.i.d. respectively according to a probability distribution over $\mathcal{X}\times\mathcal{Y}$ and a marginal distribution $\mathcal{D}_{\mathcal{X}}$. Fix $\rho > 0$ and $\kappa \in \{1, \ldots, K\}$, and consider a clustering algorithm $\mathcal{A}$ that obeys the bounded difference property with constant L and is stable. If the κ-uniformly bounded clusters found in Π_{S_u} are such that the confident level η satisfies $\eta \leq \Delta_n(\mathcal{A}_{S_u}, \mathcal{A}^\star, S_\ell)$, then for any $1 > \delta > 0$ and all $h \in \mathcal{H}$ found by the PMS$_2$L algorithm using $\mathcal{A}_{S_u}$, the following multi-class classification generalization error bound holds with probability at least $1-\delta$:*

$$R(h) \leq \widehat{R}_\rho(h) + \frac{L}{u} + \frac{2K}{\rho}(\mathfrak{R}_u^*(\mathcal{F}_{\mathcal{H}}) + \mathfrak{R}_n(\mathcal{F}_{\mathcal{H}})) + \frac{2\kappa}{\rho}\mathfrak{R}_n^*(\mathcal{F}_{\mathcal{H}}) + \frac{7G\log\frac{14G}{\delta}}{3s_*} + \sqrt{\frac{\log\frac{14}{\delta}}{t_*}} + 9\sqrt{\frac{\log\frac{14KG}{\delta}}{v_*}},$$

where $\frac{1}{s_*} \doteq \left(\frac{2G}{n-1} + \frac{G}{u-1}\right)$, $\frac{1}{t_*} \doteq \frac{L^2}{u} + \frac{1}{n}$, $\frac{1}{v_*} \doteq \frac{G\kappa u_\eta}{2u^2} + \frac{G\kappa n_\eta + K(n-n_\eta)}{2n^2}$, $n_\eta = |S_\ell \cap \mathcal{C}_\kappa(\eta)|$ and $u_\eta = |S_u \cap \mathcal{C}_\kappa(\eta)|$.

This result implies that with stable clustering algorithms obeying the bounded differences property, if the proportion of other classes than κ-predominant ones in confident clusters is less than the number of labeled examples in the exclusive disjunction of limit clusters and those found using the unlabeled training data, then with the strategy defined in algorithm PMS$_2$L we can expect to have interesting situations for learning prediction models as it is stated in the following corollary.

In this case, we can show that the convergence rate of the bound for linear classifiers is of the order

$$\tilde{O}\left(\sqrt{K/n} + K\sqrt{K/u}\right), \tag{2}$$

where, for any real valued functions f and g the equality $f(z) = \tilde{O}(g(z))$ holds, if there exists a constant $\alpha > 0$ such

that $f(z) = O(g(z)\log^{\alpha} g(z))$ [Knuth, 1976]. In the following section we present an overview of the related-work and show that in the case where the clustering technique $\mathcal{A}$ captures the true structure of the data, measured by the set of κ-uniformly bounded clusters with rate η, resulting in approximations above, then for linear kernel-based hypotheses, the convergence rate (2) is the direct extension of dimension-free convergence rates proposed in semi-supervised learning for the binary case. We also refer to [Kuznetsov *et al.*, 2014; Lei *et al.*, 2015; Maximov and Reshetova, 2016] for the state-of-the-art excess risk bounds for the supervised multi-class classification. According to these papers, the risk is linear in the number of classes and can not be further improved using the Rademacher or Gaussian complexities analysis.

As for the opposite case $n \gg u$ the pseudo-labeling step does not help to learn, and even can make the bounds worse than at the supervised case. The same situation takes place when the number of classes is comparable to the number of objects and one can not clarify whether a cluster is consistent or not. Finally, we would like to emphasize that our primary target is the most practical case with $u \gg n$ and the number of classes comparable to the number of clusters.

4 Related Works and Discussion

Semi-supervised learning (SSL) approaches exploit the geometry of data to learn a prediction function from partially labeled training sets [Seeger, 2000]. The three main SSL techniques; namely graphical, generative and discriminant approaches, were mostly developed for the binary case and tailored under smoothness, low-density separation and cluster assumptions [Zhu, 2005; Chapelle *et al.*, 2006; Amini and Usunier, 2015].

Graphical approaches construct an empirical graph where the nodes represent the training examples, and the edges of the graph reflect the similarity between them. These approaches are mostly based on label spreading algorithms that propagate the class label of each labeled node to its neighbors [Zhou *et al.*, 2003; Zhu, 2002]. Generative approaches naturally exploit the geometry of data by modeling their marginal distributions.

These methods are developed under the cluster assumption and use the Bayes rule to make a decision. In the seminal work of [Castelli and Cover, 1995] it is shown that, without extra assumptions relating marginal distribution and actual distribution of labels, a sample of unlabeled data is of (almost) no help for learning purpose. Recent work from [Ben-David *et al.*, 2008] investigated further the limitations of semi-supervised learning and concluded that theoretical results for semi-supervised learning should be accompanied by an additional assumption on the actual label distribution.

Discriminant approaches directly find the decision boundary without making any assumptions on the marginal distribution of examples. The two most popular discriminant models are without doubts co-training [Blum and Mitchell, 1998] and Transductive SVMs [Vapnik, 2000]. The co-training algorithm supposes that each observation is produced by two sources of information and that each view-specific representation is rich enough to learn the parameters of the associated classifier in the case where there are enough labeled examples available. The two classifiers are first trained separately on the labeled data. A subset of unlabeled examples is then randomly drawn and pseudo-labeled by each of the classifiers. The estimated output by the first classifier becomes the desired output for the second classifier and reciprocally. Under this setting, [Leskes, 2005] proposed a Rademacher complexity bound, where unlabeled data are used to decrease the disagreement between hypotheses from a class of functions $\mathcal{H}$ and proved that in some cases, the bound of the excess risk $|R(h) - \hat{R}(h, S_\ell)|$ for any $h \in \mathcal{H}$ is of the order $\tilde{O}\left(n^{-1/2} + u^{-1/2}\right)$.

However, transductive learning tends to produce a prediction function for only a fixed number of unlabeled examples. Transductive algorithms generally use the distribution of unsigned margins of unlabeled examples in order to guide the search of a prediction function and find the hyperplane in a feature space that separates the best l abeled examples and that does not pass through high density regions. The notion of transductive Rademacher complexity was introduced in [El-Yaniv and Pechyony, 2009]. In the best case, the excess risk bound proposed in this paper is of the order $\tilde{O}\left(u\sqrt{\min(u,n)}/(n+u)\right)$.

Our two step multi-class SSL approach is in between generative and discriminant approaches, and hence bears similarity with the study of [Urner *et al.*, 2011]. The main difference is however that the proposed approach does not rely on any pseudo-labeling mechanism and that our analyzes are based on the Rademacher complexity leading to dimension free data-dependent bounds.

On another level and under the PAC-Bayes setting, [Kääriäinen, 2005] showed that in the realizable case where the hypothesis set contains the Bayes classifier, the obtained excess risk bound takes the form $\inf_{f\in F_0} \sup_{g\in F_0} \hat{d}(f,g) + \tilde{O}\left(u^{-1/2}\right)$; where $\hat{d}(f,g)$ is a normalized empirical disagreements between two hypothesis that correctly classify the labeled set and can be of order at least $\tilde{O}\left(n^{-1/2}\right)$.

The contributions of this paper are twofold. First, we proposed a bound on the risk of a multi-class classifier trained over partially labeled training data. We derived data-dependent bounds for the generalization error of a classifier trained by minimizing an objective function that consists of an empirical risk term, estimated over the labeled training set, and a penalized term corresponding to the ratio of unlabeled examples of each cluster; within the κ bounded set of clusters, for which their predicted class does not belong to the set of the associated κ predominant classes. We support our results by extensive empirical evaluation given in [Maximov *et al.*, 2018].

Acknowledgments

This work has been partially supported by the THANATOS project funded by *Appel à projets Grenoble Innovation Recherche*. The work of YM at LANL was supported by funding from the U.S. Department of Energys Office of Electricity as part of the DOE Grid Modernization Initiative.

References

[Amini and Usunier, 2015] Massih-Reza Amini and Nicolas Usunier. *Learning with Partially Labeled and Interdependent Data*. Springer, 2015.

[Amini *et al.*, 2008] Massih-Reza Amini, François Laviolette, and Nicolas Usunier. A transductive bound for the voted classifier with an application to semi-supervised learning. In *Advances in Neural Information Processing Systems (NIPS 22)*, pages 65–72, 2008.

[Balcan and Blum, 2010] Maria-Florina Balcan and Avrim Blum. A discriminative model for semi-supervised learning. *J. ACM*, 57(3), 2010.

[Ben-David *et al.*, 2008] Shai Ben-David, Tyler Lu, and Dávid Pál. Does unlabeled data provably help? worst-case analysis of the sample complexity of semi-supervised learning. In *21st Annual Conference on Learning Theory (COLT)*, pages 33–44, 2008.

[Blum and Mitchell, 1998] Avrim Blum and Tom M. Mitchell. Combining labeled and unlabeled sata with co-training. In *11th Annual Conference on Computational Learning Theory (COLT)*, pages 92–100, 1998.

[Castelli and Cover, 1995] Vittorio Castelli and Thomas M. Cover. On the exponential value of labeled samples. *Pattern Recognition Letters*, 16(1):105–111, 1995.

[Chapelle *et al.*, 2006] Olivier Chapelle, Bernhard Schölkopf, and Alexander Zien. *Semi-supervised learning*. MIT press, 2006.

[El-Yaniv and Pechyony, 2009] Ran El-Yaniv and Dmitry Pechyony. Transductive rademacher complexity and its applications. *Journal of Artificial Intelligence Research (JAIR)*, 35:193–234, 2009.

[Kääriäinen, 2005] Matti Kääriäinen. Generalization error bounds using unlabeled data. In *Learning Theory, 18th Annual Conference on Learning Theory (COLT)*, pages 127–142, 2005.

[Knuth, 1976] Donald E. Knuth. Big omicron and big omega and big theta. *SIGACT News*, 8(2):18–24, 1976.

[Koltchinskii and Panchenko, 2002] Vladimir Koltchinskii and Dmitry Panchenko. Empirical margin distributions and bounding the generalization error of combined classifiers. *Annals of Statistics*, pages 1–50, 2002.

[Kuznetsov *et al.*, 2014] Vitaly Kuznetsov, Mehryar Mohri, and Umar Syed. Multi-class deep boosting. In *Advances in Neural Information Processing Systems*, pages 2501–2509, 2014.

[Lei *et al.*, 2015] Yunwen Lei, Urun Dogan, Alexander Binder, and Marius Kloft. Multi-class svms: From tighter data-dependent generalization bounds to novel algorithms. In *Advances in Neural Information Processing Systems*, pages 2035–2043, 2015.

[Leskes, 2005] Boaz Leskes. The value of agreement, a new boosting algorithm. In *18th Annual Conference on Learning Theory (COLT)*, pages 95–110, 2005.

[Luxburg *et al.*, 2004] Ulrike Von Luxburg, Olivier Bousquet, and Mikhail Belkin. On the convergence of spectral clustering on random samples: The normalized case. In *17th Annual Conference on Learning Theory (COLT)*, pages 457–471, 2004.

[Luxburg, 2010] Ulrike Von Luxburg. Clustering stability: An overview. *Journal Foundations and Trends in Machine Learning*, 2(3):235–274, 2010.

[Maximov and Reshetova, 2016] Yury Maximov and Daria Reshetova. Tight risk bounds for multi-class margin classifiers. *Pattern Recognition and Image Analysis*, 26(4):673–680, 2016.

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

[Rakhlin and Caponnetto, 2006] Alexander Rakhlin and Andrea Caponnetto. Stability of k-means clustering. In *Advances in Neural Information Processing Systems (NIPS 19)*, pages 1121–1128, 2006.

[Seeger, 2000] Matthias Seeger. Learning with labeled and unlabeled data. Technical report, Institute for Adaptive and Neural Computation, University of Edinburgh, 2000.

[Shamir and Tishby, 2007] Ohad Shamir and Naftali Tishby. Cluster stability for finite samples. In *Advances in Neural Information Processing Systems (NIPS 20)*, pages 1297–1304, 2007.

[Thiagarajan *et al.*, 2011] Jayaraman J. Thiagarajan, Karthikeyan Natesan Ramamurthy, and Andreas Spanias. Optimality and stability of the k-hyperline clustering algorithm. *Pattern Recognition Letters*, 32(9):1299–1304, 2011.

[Urner *et al.*, 2011] Ruth Urner, Shai Shalev-Shwartz, and Shai Ben-David. Access to unlabeled data can speed up prediction time. In *28th International Conference on Machine Learning (ICML)*, pages 641–648, 2011.

[Vapnik, 2000] Vladimir Vapnik. *The nature of statistical learning theory*. Springer Science & Business Media, 2000.

[Zhou *et al.*, 2003] Dengyong Zhou, Olivier Bousquet, Thomas Navin Lal, Jason Weston, and Bernhard Schölkopf. Learning with local and global consistency. In *Advances in Neural Information Processing Systems (NIPS 16)*, pages 321–328, 2003.

[Zhu, 2002] Xiaojin Zhu. Learning from labeled and unlabeled data with label propagation. Technical Report CMU-CALD-02-107, Carnegie Mellon University, 2002.

[Zhu, 2005] Xiaojin Zhu. Semi-supervised learning literature survey. technical report 1530. Technical report, Department of Computer Sciences, University of Wisconsin, 2005.