Journal of Artificial Intelligence Research 61 (2018) 761-786 Submitted 07/17; published 04/18

Rademacher Complexity Bounds for a Penalized Multi-class Semi-supervised Algorithm

Yury Maximov YURY@LANL.GOV
Skolkovo Institute of Science and Technology
Los Alamos National Laboratory
Theortical Division T-4 and Center for Nonlinear Studies
MS-B258, Los Alamos 87545 NM USA

Massih-Reza Amini MASSIH-REZA.AMINI@IMAG.FR
Université Grenoble Alpes
CNRS, Grenoble INP, LIG
F-38000 Grenoble France

Zaid Harchaoui ZAID@UW.EDU
University of Washington
Department of Statistics
Box 354322, Seattle, WA 98195-4322 USA

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 on different multi-class classification problems 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, Schölkopf, & Zien, 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, Laviolette, & Usunier, 2008a; El-Yaniv & Pechyony, 2009; Balcan & Blum, 2010; Urner, Shalev-Shwartz, & Ben-David, 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 findings corcerning the state-of-the-art, and finally, section 5 details experimental results that support this approach.

2. Framework and Definitions

We are interested in the study of multi-class classification problems where the output space is $\mathcal{Y} = \{1, \dots, K\}$, 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 $\kappa \in \{1, \dots, K\}$, 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

Table 1: Notations

$\mathcal{X} \subseteq \mathbb{R}^d$	Input space,
$\mathcal{Y} = \{1, \ldots, K\}$	Output space,
K (resp. G)	Number of classes (resp. clusters),
S_ℓ (resp. S_u)	The set of labeled (resp. unlabeled) training examples of size n (resp. u),
$\mathcal{A}_Z$	A clustering algorithm, $\mathcal{A}$, trained on the set Z,
$\Delta_n(\mathcal{A}_Z, \mathcal{A}_{Z'}, \tilde{Z})$	Distance between two clusterings $\mathcal{A}_Z$ and $\mathcal{A}_{Z'}$ estimated over $\tilde{Z}$ (Eq. 9),
Π_{S_u}	Partition of the unlabeled set obtained by $\mathcal{A}_{S_u}$,
$\Pi^\star$	Limit clustering of the input space obtained by a particular instantiation of $\mathcal{A}$,
$\mathcal{C}_\kappa(\eta)$	The set of κ-uniformly bounded clusters (Eq. 1),
$\mathcal{Y}_\kappa(\mathcal{C})$	κ most predominant classes found in cluster $\mathcal{C}$,
$m_h(\mathbf{x}, y)$	The margin of an example $(\mathbf{x}, y)$ over the whole set $\mathcal{Y}$ (Eq. 3),
$m'_h(\mathbf{x}, \mathcal{Y}_\kappa(\mathcal{C}))$	The margin of an unlabeled example taken with respect to $\mathcal{Y}_\kappa(\mathcal{C})$ (Eq. 7),
$\mu_h(\mathbf{x})$	The class prediction of $h \in \mathbb{R}^{\mathcal{X}\times\mathcal{Y}}$ for an example $\mathbf{x}$,
$R(h)$	Generalization error or the true risk (Eq. 2),
Ω_ρ	ρ-margin loss (Eq. 8),
$\hat{R}_\rho(h)$	Penalized empirical loss (Eq. 4),
$\Omega_\rho(h, \mathcal{C}_\kappa(\eta))$	Penalization term in $\hat{R}_\rho(h)$ estimated over $\mathcal{C}_\kappa(\eta)$ (Eq. 6),
$\widehat{R}_\rho(h, \mathcal{C}_j)$	Empirical risk defined over a single cluster $\mathcal{C}_j \in \mathcal{C}_\kappa(\eta)$ (Eq. 14).

classes such that the proportion of other classes within $\mathcal{C}$ not belonging to $\mathcal{Y}_\kappa(\mathcal{C})$ is less than η/G :

$$\mathcal{C}_\kappa(\eta) = \left\{ \mathcal{C} \in \Pi_{S_u} : P_n((\mathbf{x}, y) \in \mathcal{C} \wedge y \in \mathcal{Y} \setminus \mathcal{Y}_\kappa(\mathcal{C})) \leq \frac{\eta}{G} \right\}. \tag{1}$$

Where P_n the uniform probability distribution over S_ℓ; is defined for any subset $B \subseteq S_\ell$, as $P_n(B) = \frac{1}{n}\text{card}(B)$.

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:

$$R(h) = \mathbb{E}_{(\mathbf{x},y)\sim\mathcal{D}} \left[\mathbb{1}_{m_h(\mathbf{x},y)\leqslant 0} \right], \tag{2}$$

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

$$m_h(\mathbf{x}, y) = h(\mathbf{x}, y) - \max_{y' \in \mathcal{Y}\setminus\{y\}} h(\mathbf{x}, y'). \tag{3}$$

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

$$\widehat{R}_\rho(h) = \widehat{R}_\rho(h, S_\ell) + \Omega_\rho(h, \mathcal{C}_\kappa(\eta)), \tag{4}$$

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

$$\widehat{R}_\rho(h, S_\ell) = \frac{1}{n} \sum_{(\mathbf{x},y)\in S_\ell} \Phi_\rho(m_h(\mathbf{x}, y)), \tag{5}$$

Algorithm 1: Pseudo-code of the PMS$_2$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, $\mathcal{A}_{S_u} : \mathcal{X} \to \{1,\dots,G\}$ 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 Eq. (1)

Stage 2: Find a hypothesis $h^* \in \mathcal{H}$ that minimizes the penalized objective function (Eq. 4) :

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

Output: h^*

---

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 $\mathcal{C}_\kappa(\eta)$;

$$\Omega_\rho(h, \mathcal{C}_\kappa(\eta)) = \frac{1}{u}\sum_{\mathcal{C}\in\mathcal{C}_\kappa(\eta)}\sum_{\mathbf{x}\in\mathcal{C}} \Phi_\rho(m'_h(\mathbf{x}, \mathcal{Y}_\kappa(\mathcal{C}))), \tag{6}$$

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, $\mathcal{Y}_\kappa(\mathcal{C})$:

$$m'_h(\mathbf{x}, \mathcal{Y}_\kappa(\mathcal{C})) = \max_{y\in\mathcal{Y}_\kappa(\mathcal{C})} h(\mathbf{x}, y) - \max_{y\in\mathcal{Y}\setminus\mathcal{Y}_\kappa(\mathcal{C})} h(\mathbf{x}, y),\ \mathbf{x}\in\mathcal{C}\subset\mathcal{C}_\kappa(\eta), \tag{7}$$

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

$$\forall z\in\mathbb{R}, \Phi_\rho(z) = \begin{cases} 0 & \text{if } \rho \geqslant z, \\ 1 - z/\rho & \text{if } 0 < z < \rho, \\ 1 & \text{if } z \leqslant 0. \end{cases} \tag{8}$$

Table 1 summarizes notations used throughout the paper and the pseudo-code of the proposed 2-step approach, referred to as Penalized Multi-Class Semi-Supervised Learning (PMS$_2$L) in the following, is given in algorithm 1.

The algorithm shares similarities with algorithms proposed by Amini, Truong, and Goutte (2008b) and Urner et al. (2011), where the k-NN technique was used to increase the size of the labeled training data by pseudo-labeling unlabeled examples that are in the nearest neighborhood of labeled examples, for binary classification and bipartite ranking. In the work of Rigollet (2007), another two-step semi-supervised procedure is proposed, where in the first stage a clustering of the feature space derived from the unlabeled data is produced and then each unlabeled observation, in a given cluster is assigned the same class label than the majority of labeled examples belonging to that class within the cluster.

In the present work, we tackle a more general situation by considering multi-class classification problems and by relaxing the pseudo-labeling part which may be too aggressive in the multi-class case. Our analysis is based on the ability of a clustering technique to capture the structure of the data, and the ability of the classifier to identify predominant classes in κ-uniformly bounded clusters, leading to a multi-class definition of the cluster assumption which states that penalization over κ-uniformly bounded clusters with a bounded confident level η helps learning.

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 $\mathcal{A}_.$; 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 $\mathcal{A}_.$ over two sets Z and Z', and defined as :

$$\Delta_n\left(\mathcal{A}_Z, \mathcal{A}_{Z'}, \tilde{Z}\right) = \min_{\pi}\left[\frac{1}{n}\sum_{\mathbf{x}\in\tilde{Z}} \mathbb{1}_{\mathcal{A}_Z(\mathbf{x})\neq\pi(\mathcal{A}_{Z'}(\mathbf{x}))}\right], \tag{9}$$

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 (see Th. 6 in the Appendix). Hence, the clustering algorithm $\mathcal{A}_.$ 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 :

$$\Delta(\mathcal{A}_Z, \mathcal{A}_{Z'}) = \mathbb{E}_{\tilde{Z}\sim\mathcal{D}_{\mathcal{X}}^n}\left[\Delta_n\left(\mathcal{A}_Z, \mathcal{A}_{Z'}, \tilde{Z}\right)\right] \leqslant \frac{L}{|Z|}. \tag{10}$$

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, Bousquet, and Belkin (2004), Rakhlin and Caponnetto (2006) and Thiagarajan, Ramamurthy, and Spanias (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 :

$$\mathbb{E}_{Z\sim\mathcal{D}_{\mathcal{X}}^{|Z|}}\left[\Delta(\mathcal{A}_Z, \mathcal{A}^\star)\right] \leqslant \frac{L}{|Z|}. \tag{11}$$

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 :*

$$\Delta_n(\mathcal{A}_{S_u}, \mathcal{A}^\star, S_\ell) \leqslant \frac{L}{u} + L\sqrt{\frac{\log\frac{2}{\delta}}{2u}} + \sqrt{\frac{\log\frac{2}{\delta}}{2n}}. \tag{12}$$

The proof is given in Appendix B. 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 $\texttt{PMS}_2\texttt{L}$ becomes consistent. This result is stated in Theorem (3) which provides bounds on the generalization error of a multi-class classifier trained with the penalized empirical loss defined above (Eq. 4).

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:

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

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

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

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 of the theorem is based on the following Lemma that provides generalization bounds over the true risk of any classifier h, found by algorithm $\texttt{PMS}_2\texttt{L}$ and estimated within a single confident cluster; $\mathcal{C}_j\in\mathcal{C}_\kappa(\eta)\subseteq\Pi_{S_u}$:

$$R(h,\mathcal{C}_j) = \mathbb{E}[\mu_h(\mathbf{x})\neq y \wedge \mathbf{x}\in\mathcal{C}_j], \tag{13}$$

with respect to the estimated empirical risk :

$$\widehat{R}_\rho(h,\mathcal{C}_j) = \frac{1}{n}\sum_{(\mathbf{x},y)\in S_\ell\cap\mathcal{C}_j} \Phi_\rho(m_h(\mathbf{x},y)) + \frac{1}{u}\sum_{\mathbf{x}\in S_u\cap\mathcal{C}_j} \Phi_\rho(m'_h(\mathbf{x},\mathcal{Y}_\kappa(\mathcal{C}_j))). \tag{14}$$

Lemma 2 *Let $\mathcal{H} \subseteq \mathbb{R}^{\mathcal{X}\times\mathcal{Y}}$ be a hypothesis set where $\mathcal{Y} = \{1,\dots,K\}$, and 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 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 $\rho > 0$, $\kappa \in \{1,\dots,K\}$ then for any $1 > \delta > 0$, the following multi-class classification generalization error bound holds with probability at least $1-\delta$ for all $h \in \mathcal{H}$ learned by algorithm 1 over a single κ-uniformly bounded cluster $\mathcal{C}_j \in \mathcal{C}_\kappa(\eta)$ derived from S_u by a clustering algorithm $\mathcal{A}_{S_u}$ that partitions the input space into G clusters :*

$$R(h,\mathcal{C}_j) \leq \widehat{R}_\rho(h,\mathcal{C}_j) + \frac{\eta}{G} + \frac{2\kappa}{\rho}\mathfrak{R}^*_{n,j}(\mathcal{F}_\mathcal{H}) + \frac{2K}{\rho}\mathfrak{R}^*_{u,j}(\mathcal{F}_\mathcal{H}) + 5\sqrt{\frac{\kappa n_\eta(j)\log\frac{16K}{\delta}}{2n^2}} + 5\sqrt{\frac{\kappa u_\eta(j)\log\frac{16K}{\delta}}{2u^2}} + \frac{7\log\frac{8}{\delta}}{3(n-1)} + \frac{7\log\frac{8}{\delta}}{3(u-1)},$$

where $n_\eta(j) = |S_\ell \cap \mathcal{C}_j|$, $u_\eta(j) = |S_u \cap \mathcal{C}_j|$, $\mathfrak{R}^*_{n,j} = \mathbb{E}_\sigma \sup_{f\in\mathcal{F}_\mathcal{H}} \frac{2}{n}\sum_{\mathbf{x}_i\in S_\ell\cap\mathcal{C}_j}\sigma_i f(\mathbf{x}_i)$, and $\mathfrak{R}^*_{u,j} = \mathbb{E}_\sigma \sup_{f\in\mathcal{F}_\mathcal{H}} \frac{2}{u}\sum_{\mathbf{x}_i\in S_u\cap\mathcal{C}_j}\sigma_i f(\mathbf{x}_i)$.

The proof is provided in Appendix B. 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 $\texttt{PMS}_2\texttt{L}$ as stated below.

Theorem 3 *Let $\mathcal{H} \subseteq \mathbb{R}^{\mathcal{X}\times\mathcal{Y}}$ be a hypothesis set where $\mathcal{Y} = \{1,\dots,K\}$, 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 $\kappa \in \{1,\dots,K\}$, 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 $\texttt{PMS}_2\texttt{L}$ 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 $u_\eta = |S_u \cap \mathcal{C}_\kappa(\eta)|$.

The proof is provided in Appendix B. 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 $\texttt{PMS}_2\texttt{L}$ we can expect to have interesting situations for learning prediction models as it is stated in the following corollary.

Consider kernel-based hypotheses with $\mathfrak{K} : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ a PSD kernel and $\Phi : \mathcal{X} \to \mathbb{H}$ its associated feature mapping function, defined as :

$$\mathcal{H}_B = \left\{ (\mathbf{x}, y) \in \mathcal{X} \times \mathcal{Y} \mapsto \langle \Phi(\mathbf{x}), \mathbf{w}_y \rangle \mid \mathbf{W} = (\mathbf{w}_1, \dots, \mathbf{w}_K), \|\mathbf{W}\|_{\mathbb{H},2} \leq B \right\}.$$

Where $\|\mathbf{W}\|_{\mathbb{H},2}$ is the Frobenius norm of the parameter matrix for a linear kernel, or the $L_{\mathbb{H},2}$ group norm of $\mathbf{W}$, defined as

$$\|\mathbf{W}\|_{\mathbb{H},2} = \sqrt{\sum_{k=1}^{K} \|\mathbf{w}_k\|_{\mathbb{H}}^2}.$$

In this case, we can derive the following corollary from Theorem 3 :

Corollary 4 *Let $\mathfrak{K} : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ be a PSD kernel and let $\Phi : \mathcal{X} \to \mathbb{H}$ be the associated feature mapping function. Assume that there exists $R > 0$ such that $\mathfrak{K}(\mathbf{x}, \mathbf{x}) \leq R^2$ for all $\mathbf{x} \in \mathcal{X}$. Then for any $1 > \delta > 0$ and under the conditions and the definitions of Theorem 3, the following multi-class classification error bound holds for all hypothesis $h \in \mathcal{H}_B$ learned by the proposed algorithm over the set of κ-uniformly bounded set of clusters, $\mathcal{C}_\kappa(\eta)$, with probability at least $1 - \delta$:*

$$R(h) \leq \widehat{R}_\rho(h) + \frac{L}{u} + \frac{2}{\rho} RB \sqrt{\frac{3k_*^2}{s_*}} + \frac{7 \log \frac{14G}{\delta}}{3s_*} + \sqrt{\frac{\log \frac{14}{\delta}}{t_*}} + 5\sqrt{\frac{3 \log \frac{14KG}{\delta}}{v_*}}$$

where $\frac{1}{s_*} \doteq \left(\frac{2G}{n-1} + \frac{G}{u-1} \right)$, $\frac{1}{t_*} \doteq \frac{L^2}{u} + \frac{1}{n}$, $\frac{1}{v_*} \doteq \frac{G\kappa u_\eta}{2u^2} + \frac{G\kappa n_\eta}{2n^2} + \frac{K(n-n_\eta)}{2n^2}$ and $\frac{k_*^2}{v_*} \doteq K^2 \frac{Gu_\eta}{u^2} + \kappa^2 \frac{Gn_\eta}{n^2} + K^2 \frac{n-n_\eta}{n^2}$.

Proof. From the proposition (8.1) in (Mohri, Rostamizadeh, & Talwalkar, 2012), and the Cauchy-Schwartz inequality $\left(\sum_{j=1}^{G} a_j b_j \right)^2 \leqslant \left(\sum_{j=1}^{G} a_j^2 \right) \left(\sum_{j=1}^{G} b_j^2 \right)$ with $b_j = 1$ and $a_j = \sqrt{u_\eta(j)}, \forall j$; the Rademacher complexity of the class of linear classifiers in the feature space can be bounded as :

$$\mathfrak{R}_u^*(\mathcal{F}_\mathcal{H}) \leqslant \sum_{\mathcal{C}_j \in \mathcal{C}_\kappa(\eta)} \frac{2}{u} RB \sqrt{u_\eta(j)} \leqslant 2RB \sqrt{\frac{Gu_\eta}{u^2}},$$

where $u_\eta(j)$ in the number of unlabeled examples in η-confident cluster $\mathcal{C}_j$ and $u_\eta = \sum_j u_\eta(j)$ is the total number of unlabeled examples within a set of confident clusters $\mathcal{C}_\kappa(\eta)$.

Similarly, if $n_\eta(j)$ is the number of unlabeled examples in $\mathcal{C}_j \in \mathcal{C}_\kappa(\eta)$ we have :

$$\mathfrak{R}_n^*(\mathcal{F}_\mathcal{H}) \leqslant \sum_{\mathcal{C}_j \in \mathcal{C}_\kappa(\eta)} \frac{2}{n} RB \sqrt{n_\eta(j)} \leqslant 2RB \sqrt{\frac{Gn_\eta}{n^2}},$$

and also $\mathfrak{R}_n(\mathcal{F}_\mathcal{H}) \leqslant 2RB\sqrt{\frac{n-n_\eta}{n^2}}$. Applying the Cauchy-Schwartz inequality again we finally get :

$$2K\mathfrak{R}_n(\mathcal{F}_\mathcal{H}) + 2\kappa\mathfrak{R}_n^*(\mathcal{F}_\mathcal{H}) + 2K\mathfrak{R}_u^*(\mathcal{F}_\mathcal{H}) \leqslant 2RB\sqrt{\frac{3k_*^2}{s_*}}$$

$\square$

The non-empirical terms of this bound determine the convergence rate of the proposed penalized semi-supervised mutli-class algorithm, and hence following (Vapnik, 2000, Thm. 2.1, p.38), gives insights on its consistency. These terms may be better explained using orders of magnitude (Knuth, 1976). If we now consider the common situation in semi-supervised learning where $u \gg n$, and $n_\eta \approx n, u_\eta \approx u$, and $\kappa = O(1)$, $L = O(1)$, $G = O(K)$ then

$$\frac{k_*^2}{v_*} \doteq K^2 \frac{G u_\eta}{u^2} + \kappa^2 \frac{G n_\eta}{n^2} + K^2 \frac{n - n_\eta}{n^2} = O\left(\frac{K^3}{u} + \frac{K}{n}\right),$$

$$\frac{1}{v_*} \doteq \frac{G\kappa u_\eta}{u^2} + \frac{G\kappa n_\eta}{n^2} + \frac{K(n - n_\eta)}{n^2} = O\left(\frac{K}{u} + \frac{K}{n}\right),$$

and

$$\frac{1}{s_*} = O\left(\frac{K}{n} + \frac{K}{u}\right), \ \ \frac{1}{t_*} \doteq \frac{L^2}{u} + \frac{1}{n} = O\left(\frac{1}{u} + \frac{1}{n}\right).$$

The convergence rate of the bound of corollary 4 is of the order

$$\tilde{O}\left(\sqrt{\frac{K}{n}} + K\sqrt{\frac{K}{u}}\right), \tag{15}$$

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 (15) is the direct extension of dimension-free convergence rates proposed in semi-supervised learning for the binary case.

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 & 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, Bousquet, Lal, Weston, & Schölkopf, 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, Lu, and Pál (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 & 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 $\tilde{O}\left(n^{-1/2} + u^{-1/2}\right)$. Another study in this line of research is that of Tolstikhin, Zhivotovskiy, and Blanchard (2015). 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 labeled 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

Table 2: Summary of the convergence rates of dimension free bounds of excess risks for different SSL approaches.

Order of convergence rates	Case; Reference
$\tilde{O}\left(\frac{u\sqrt{\min(u,n)}}{n+u}\right)$	Binary; (El-Yaniv & Pechyony, 2009)
$\tilde{O}\left(\frac{1}{n} + \frac{1}{\sqrt{u}}\right)$	Binary;(Balcan & Blum, 2010)
$\tilde{O}\left(\frac{1}{\sqrt{n}} + \frac{1}{\sqrt{u}}\right)$	Binary; (Leskes, 2005)
$\tilde{O}\left(\frac{1}{\sqrt{n}} + \frac{1}{\sqrt{u}}\right)$	Binary; (Kääriäinen, 2005)
$\tilde{O}\left(\frac{\sqrt{K}}{\sqrt{n}} + \frac{K^{3/2}}{\sqrt{u}}\right)$	Multi-class; Corollary 4

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 convergence rates of the mentioned bounds are sum up in Table 2. From these results, it becomes apparent that the convergence rate deduced from corollary 4, (Equation 15) extends those found by Kääriäinen (2005) and Leskes (2005) for multi-class classification.

5. Experimental Results

We perform experiments on six publicly available datasets. The three first ones are `Fungus`, `Birds` and `Athletics` that consist of three aggregations of lead nodes that go down from parent nodes in the ImageNet hierarchy (Russakovsky, Deng, Su, Krause, Satheesh, Ma, Huang, Karpathy, Khosla, Bernstein, Berg, & Fei-Fei, 2015). Each image is characterized by a Fisher vector representation as described by Harchaoui, Douze, Paulin, Dudík, and Malick (2012). The three others collections are respectively the `MNIST` database of handwritten digits, the pre-processed 20 Newsgroups (`20-NG`) collection (Chang & Lin, 2011) and the `USPS` dataset (Hastie, Tibshirani, & Friedman, 2009). Table 2 resumes the characteristics of these datasets. The proportions of training and test sets were kept fixed to those given in the released data files. Within the training set ($S_\ell \cup S_u$) we randomly sampled labeled examples S_ℓ, with different sizes, and used the remaining as unlabeled data.

Table 2: Characteristics of datasets used in our experiments.

dataset	$\lvert S_\ell \cup S_u\rvert$	size of the test	dimension, d	# of classes, K
`Birds`	5785	5596	4096	196
`Athletics`	28752	28727	4096	51
`Fungus`	50270	50271	4096	134
`20-NG`	15936	3393	62061	20
`MNIST`	60000	10000	780	10
`USPS`	7291	2007	256	10

To validate the proposed penalized based multi-class semi-supervised learning approach ($\texttt{PMS}_2\texttt{L}$), we compared its results with respect to a multi-class extension of a popular SSL algorithm proposed within each of the Generative, Graphical and Discriminant approaches. More precisely we considered the extension of the label propagation algorithm to the multi-class case (`McLP`) proposed by Wang, Tu, and Tsotsos (2013). A generative SSL model based on the mixture of Gaussians ($\texttt{S}_2\texttt{GM}$), the extension of TSVM (Joachims, 1999) to the multi-class case (`McTSVM`), and a purely supervised technique which does not make use of any unlabeled examples in the training stage (`SUP`).

As the clustering algorithm $\mathcal{A}$, we employed the Nearest Neighbor Clustering technique proposed by Bubeck and Luxburg (2009), and fixed $m = 4K$, $\kappa = 2$ and $\eta = 10^{-3}$. Meaning that each cluster in $\mathcal{C}_\kappa(\eta)$ is mainly composed of the two most predominant classes within it. For the second stage of $\texttt{PMS}_2\texttt{L}$, as well as for `SUP` and `McTSVM`, we adapted the aggregated one-versus-all approach using a linear kernel `SVM` that respects the conditions of corollary 4. The penalized objective function can be easily implemented using convex optimization tools for convex surrogates of the 0/1 loss. The parameter C of the `SVM` classifier is determined by five fold cross-validation in logarithmic range between 10^{-4} and 10^4 over the available labeled training data. Results are evaluated over the test set using the accuracy, and the reported performance is averaged over 25 random (labeled/unlabeled/test) sets of the initial collections.

Table 3 summarizes results obtained by `SUP`, $\texttt{PMS}_2\texttt{L}$, `McLP`, $\texttt{S}_2\texttt{GM}$ and `McTSVM` when a very small proportion of labeled training data is used in the learning of the models. We use boldface to indicate the highest performance rates, and the symbol $^\downarrow$ indicates that performance is significantly worse than the best result, according to a Wilcoxon rank sum test used at a p-value threshold of 0.05 (Lehmann, 1975). From these results it becomes clear that

- The algorithm $\texttt{PMS}_2\texttt{L}$ performs significantly better than all of the four other algorithms, and it improves over `SUP` by an average of 1.5 to 6.5% on different datasets.

- `McLP` and `McTSVM` also perform better than `SUP`, though not in the same range than previously, while the mixture of Gaussians $\texttt{S}_2\texttt{GM}$ does worse than `SUP` especially in the cases where the dimension of the problem is high.

- Finally, the difference in performance between $\texttt{PMS}_2\texttt{L}$ and `McTSVM` is smaller than the one between the former and `McLP`.

Our analysis of these results is that the Nearest Neighbor Clustering technique (Bubeck & Luxburg, 2009) is effectively able to map correctly the considered data, into homogenous clusters containing mostly unlabeled examples of the same class than the $\kappa = 2$ most predominant classes contained in them. In this case, the penalized term of the objective function used to learn the classifier (Equation 4) forcefully helps to pick a better hypothesis in the set of linear classifiers, than when only labeled training data are used. Hence, for unlabeled examples within a given cluster, the constraint of predicting the same classes than the $\kappa = 2$ most predominant classes of that cluster makes the decision boundary to pass through regions where the unsigned margins of unlabeled

Table 3: Means and standard deviations of the classification accuracy on test data over the 25 trials for each data set. n_y refers to the average number of labeled examples per class in each data set. $^\downarrow$ indicates statistically significantly worse performance than the best result, shown in bold, according to a Wilcoxon rank sum test ($p < 0.05$) (Lehmann, 1975).

Dataset	n_y	$n/(n+u)$	SUP	$\texttt{PMS}_2\texttt{L}$	McLP	$\texttt{S}_2\texttt{GM}$	McTSVM
Birds	5	0.18	$.294^\downarrow \pm .03$	$\mathbf{.344} \pm .03$	$.303^\downarrow \pm .06$	$.286^\downarrow \pm .08$	$.312^\downarrow \pm .04$
Athletics	43	0.08	$.258^\downarrow \pm .03$	$\mathbf{.273} \pm .02$	$.259^\downarrow \pm .05$	$.246^\downarrow \pm .07$	$.263 \pm .04$
Fungus	15	0.04	$.121^\downarrow \pm .03$	$\mathbf{.160} \pm .03$	$.125^\downarrow \pm .06$	$.107^\downarrow \pm .05$	$.134^\downarrow \pm .04$
20-NG	16	0.02	$.468^\downarrow \pm .05$	$\mathbf{.531} \pm .03$	$.476^\downarrow \pm .06$	$.452^\downarrow \pm .04$	$.484^\downarrow \pm .04$
MNIST	120	0.02	$.767^\downarrow \pm .03$	$\mathbf{.799} \pm .02$	$.771^\downarrow \pm .05$	$.758^\downarrow \pm .06$	$.781^\downarrow \pm .01$
USPS	14	0.02	$.790^\downarrow \pm .03$	$\mathbf{.821} \pm .02$	$.796^\downarrow \pm .04$	$.788^\downarrow \pm .06$	$.801^\downarrow \pm .02$

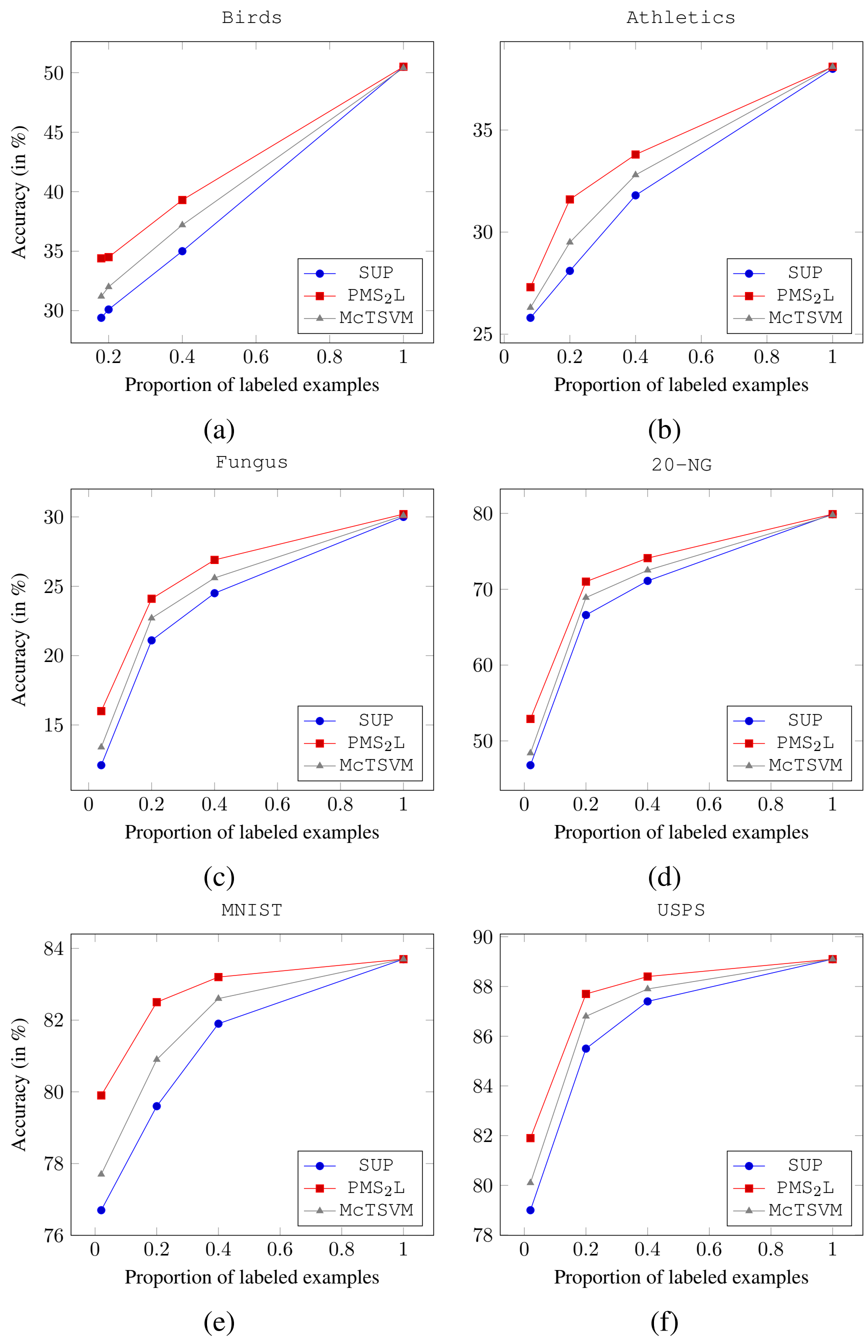

Figure 1: Accuracy in percentage with respect to the proportion of labeled examples in the initial training set for ImageNet `Birds` (a), `Athletics` (b), `Fungus` (c); `20-NG` (d), `MNIST` (e), and `USPS` (f). Each reported performance on the test is averaged over 25 random (labeled/unlabeled/test) sets of the initial collections.

examples are small. As stated in section 4, this is exactly how TSVM works, and the proximity of results between McTSVM and PMS_2L, compared to the two other SSL algorithms can be explained by the similitude of the assumptions leading to the development of these models.

However, the fundamental difference between these two algorithms in the iterative pseudo-labeling of unlabeled examples (or not), would do that, when the proportion of labeled training data is small, the iterative pseudo-labeling steps of McTSVM injects noise into the learning process at the same level or even more than the true labeled information. The question therefore arises as to how these two techniques behave for more labeled training data available at the learning phase?

In order to analyze more finely this situation, we compared SUP, PMS_2L and McTSVM for an increasing size of the labeled training data. Figure 5, illustrates this by showing the accuracy (in percentage) with respect to the number of labeled examples in the initial labeled training set S_ℓ. The main observations drawn from these results, are:

- As expected, all performance curves increase monotonously with respect to the additional labeled data and converge to the same performance. We note that when all the labeled training data are used for learning the linear SVM gives the same results than those reported in the state-of-the art; e.g. the MLP model with no hidden layer on USPS (LeCun, Bottou, Bengio, & Haffner, 2001) and (Maji & Malik, 2009).

- Though McTSVM takes advantage of unlabeled data in its learning process, it is outperformed by PMS_2L.

- On ImageNet Birds and MNIST, a non-negligible quantity of labeled examples is necessary for SUP to catch the performance of PMS_2L learned with the same proportion of labeled data than the one of Table 3, and the remaining unlabeled training data.

These behaviour first suggest that when enough labeled data is available, unlabeled data do not serve the learning algorithm as for the reverse situation. These results suggest that for SSL discriminant techniques designed following the low density separation hypothesis, a more convenient approach than the pseudo-labeling strategy, used in most of these techniques, would be the incorporation of a penalized factor concerning unlabeled examples into the objective of the learning algorithm as the one proposed in Equation 4.

6. Conclusion

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. The analysis of this bound for kernel-based hypotheses reveals a convergence rate that is an extension to the multi-class case, of some other rates over the bounds of the excess risk proposed in the literature. Empirical results on a various datasets support our findings by showing that the proposed algorithm is competitive compared to different extensions of binary semi-supervised learning algorithms and that it may significantly increase classification performance in the most interesting situation, when there are few labeled data available for training.

Acknowledgments

The authors are thankful to the anonymous reviewers for their numerous helpful suggestions which significantly improved the paper. 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 funded by DOE/GMLC 2.0 project: "Emergency Monitoring and controls through new technologies and analytics".

Appendix A. Mathematical Tools

Theorem 5 (McDiarmid's (1989) inequality) *Let $X_1, \dots, X_u \in \mathcal{X}^u$ be a set of $u \geqslant 1$ independent random variables and assume that there exist $c_1, \dots, c_u > 0$ such that $\phi : \mathcal{X}^u \to \mathbb{R}$ satisfies the following condition:*

$$|\phi(x_1, \dots, x_i, \dots, x_u) - \phi(x_1, \dots, x_i^{'}, \dots, x_u)| \leq c_i,$$

for all $i \in [\![1, u]\!]$ and any points $x_1, \dots, x_u, x_i^{'} \in \mathcal{X}$. Let $\phi(S)$ denote $\phi(X_1, \dots, X_u)$, then, for all $\epsilon > 0$, the following inequalities hold:

$$\mathbb{P}[\phi(S) - \mathbb{E}[\phi(S)] \geq \epsilon] \leq \exp\left(\frac{-2\epsilon^2}{\sum_{i=1}^u c_i^2}\right) \text{ and } \mathbb{P}[\phi(S) - \mathbb{E}[\phi(S)] \leq \epsilon] \quad \leq \exp\left(\frac{-2\epsilon^2}{\sum_{i=1}^u c_i^2}\right).$$

Theorem 6 (Minimal matching distance) *Let $\mathcal{A}_{Z_1}$ and $\mathcal{A}_{Z_2}$ be two partitions obtained by a clustering algorithm $\mathcal{A}$ over two finite sets Z_1 and Z_2. Then for any sample set $\tilde{Z} \subseteq \mathcal{X}$, of size n, where $\forall \mathbf{x} \in \tilde{Z}, \mathcal{A}_Z(\mathbf{x}) \in \{1, \dots, G\}$ is the partition of $\mathbf{x}$; the function*

$$\Delta_n : (\mathcal{A}_{Z_1}, \mathcal{A}_{Z_2}, \tilde{Z}) \mapsto \min_{\pi} \frac{1}{n} \sum_{\mathbf{x} \in \tilde{Z}} \mathbb{1}_{\mathcal{A}_{Z_1}(\mathbf{x}) \neq \pi(\mathcal{A}_{Z_2}(\mathbf{x}))},$$

is a metric over the space of clusterings.

Proof. For all $\mathcal{A}_{Z_1}, \mathcal{A}_{Z_2}, \mathcal{A}_{Z_3}$, and $\tilde{Z}$ the following conditions are indeed satisfied :

1. *non-negativity:* $\Delta_n(\mathcal{A}_{Z_1}, \mathcal{A}_{Z_2}, \tilde{Z}) \geq 0$,
2. *identity:* $\Delta_n(\mathcal{A}_{Z_1}, \mathcal{A}_{Z_2}, \tilde{Z}) = 0 \Leftrightarrow \mathcal{A}_{Z_1} = \mathcal{A}_{Z_2}$,
3. *symmetry:* $\Delta_n(\mathcal{A}_{Z_1}, \mathcal{A}_{Z_2}, \tilde{Z}) = \Delta_n(\mathcal{A}_{Z_2}, \mathcal{A}_{Z_1}, \tilde{Z})$,
4. *triangle inequality:* $\Delta_n(\mathcal{A}_{Z_1}, \mathcal{A}_{Z_2}, \tilde{Z}) \leq \Delta_n(\mathcal{A}_{Z_1}, \mathcal{A}_{Z_3}, \tilde{Z}) + \Delta_n(\mathcal{A}_{Z_2}, \mathcal{A}_{Z_3}, \tilde{Z})$.

The last inequality is due to the fact that for any permutations π, π_1 and π_2, we have :

$$\forall \mathbf{x} \in \tilde{Z}, \mathbb{1}_{\mathcal{A}_{Z_1}(\mathbf{x}) \neq \pi(\mathcal{A}_{Z_2}(\mathbf{x}))} \leqslant \mathbb{1}_{\mathcal{A}_{Z_1}(\mathbf{x}) \neq \pi_1(\mathcal{A}_{Z_3}(\mathbf{x}))} + \mathbb{1}_{\mathcal{A}_{Z_3}(\mathbf{x}) \neq \pi_2(\mathcal{A}_{Z_2}(\mathbf{x}))},$$

summing over all $\mathbf{x} \in \tilde{Z}$ gives:

$$\frac{1}{n} \sum_{\mathbf{x} \in \tilde{Z}} \mathbb{1}_{\mathcal{A}_{Z_1}(\mathbf{x}) \neq \pi(\mathcal{A}_{Z_2}(\mathbf{x}))} \leqslant \frac{1}{n} \sum_{\mathbf{x} \in \tilde{Z}} \mathbb{1}_{\mathcal{A}_{Z_1}(\mathbf{x}) \neq \pi_1(\mathcal{A}_{Z_3}(\mathbf{x}))} + \frac{1}{n} \sum_{\mathbf{x} \in \tilde{Z}} \mathbb{1}_{\mathcal{A}_{Z_3}(\mathbf{x}) \neq \pi_2(\mathcal{A}_{Z_2}(\mathbf{x}))}.$$

As the last inequality is valid for any permutations π, π_1 and π_2 over $\tilde{Z}$ we have :

$$\begin{aligned}
\Delta_n(\mathcal{A}_{Z_1}, \mathcal{A}_{Z_2}, \tilde{Z}) &= \min_{\pi} \frac{1}{n} \sum_{\mathbf{x} \in \tilde{Z}} \mathbb{1}_{\mathcal{A}_{Z_1}(\mathbf{x}) \neq \pi(\mathcal{A}_{Z_2}(\mathbf{x}))} \\
&\leqslant \min_{\pi_1} \frac{1}{n} \sum_{\mathbf{x} \in \tilde{Z}} \mathbb{1}_{\mathcal{A}_{Z_1}(\mathbf{x}) \neq \pi_1(\mathcal{A}_{Z_3}(\mathbf{x}))} + \min_{\pi_2} \frac{1}{n} \sum_{\mathbf{x} \in \tilde{Z}} \mathbb{1}_{\mathcal{A}_{Z_3}(\mathbf{x}) \neq \pi_2(\mathcal{A}_{Z_2}(\mathbf{x}))} \\
&= \Delta_n(\mathcal{A}_{Z_1}, \mathcal{A}_{Z_3}, \tilde{Z}) + \Delta_n(\mathcal{A}_{Z_2}, \mathcal{A}_{Z_3}, \tilde{Z}).
\end{aligned}$$

□

Theorem 7 (Data-dependent Bennett's inequality, see Maurer & Pontil, 2009, Thm. 4) *Let X, $X_1, \ldots, X_n$ be i.i.d. random variables with values in $[0,1]$ and let $\delta > 0$. Then with probability at least $1-\delta$ in $(X_1, \ldots, X_n)$ we have*

$$\mathbb{E}\left[X\right] - \frac{1}{n}\sum_{i=1}^{n} X_i \leqslant \sqrt{\frac{2V_n(X)\log\frac{2}{\delta}}{n}} + \frac{7\log\frac{2}{\delta}}{3(n-1)},$$

where $V_n(X)$ is the sample variance

$$V_n(X) = \frac{1}{n(n-1)} \sum_{1 \leqslant i < j \leqslant n} (X_i - X_j)^2$$

Lemma 8 (Lemma 8.1 see Mohri et al., 2012) *Let $\mathcal{F}_1, \ldots, F_l$ be l hypothesis sets in $\mathbb{R}^{\mathcal{X}}$, $l \geq 1$, and let $\mathcal{G} = \{\max(h_1, \ldots, h_l) : h_i \in \mathcal{F}_i\}$, $1 \leq i \leq l$. Then, for any sample S of size n, the empirical Rademacher complexity of G can be upper bounded as follows:*

$$\mathfrak{R}_n^*(\mathcal{G}) \leqslant \sum_{i=1}^{l} \mathfrak{R}_n^*(\mathcal{F}_i).$$

Theorem 9 (Rademacher generalization bounds see Mohri et al., 2012, Thm. 8.1) *Let G be a family of functions mapping from $\mathcal{X}$ to $[0,1]$. Then for any $1 > \delta > 0$, with probability at least $1-\delta$ we have for all $g \in G$:*

$$\mathbb{E}[g] \leqslant \frac{1}{n}\sum_{i=1}^{n} g(z_i) + \mathfrak{R}_n^*(G) + 3\sqrt{\frac{\log\frac{2}{\delta}}{2n}}$$

Definition 10 (L-regular loss, see Lei, Dogan, Binder, & Kloft, 2015, definition 2) *A loss function ℓ is said to be L-regular if :*

1. *$\ell(t)$ bounds the 0-1 loss from above: $\ell(t) \geqslant 1_{t \leqslant 0}$;*
2. *ℓ is L-Lipschitz in the sense $|\ell(t_1) - \ell(t_2)| \leqslant L|t_1 - t_2|$;*
3. *$\ell(t)$ is decreasing and it has a zero point c_ℓ, i.e., $\ell(c_\ell) = 0$.*

Theorem 11 (Multi-class Rademacher generalization bounds; see Lei et al., 2015, remark 6) *Let $\mathcal{F}_{\mathcal{H}} \subset \mathbb{R}^{\mathcal{X}\times\mathcal{Y}}$ be a hypothesis class with $\mathcal{Y} = \{1, \dots, K\}$. Let ℓ be a L-regular loss function and denote $B_\ell \doteq \sup_{(\mathbf{x},y),h} \ell(m_h(\mathbf{x}, y))$.*

Suppose that the examples $S_\ell = \{(\mathbf{x}_i, y_i); i \in \{1, \dots, n\}\}$ are i.i.d with respect to a fixed yet unknown probability distribution defined on $\mathcal{X} \times \mathcal{Y}$. Then, for any $\delta > 0$, with probability at least $1 - \delta$, the following multi-class classification generalization bound holds for any $h \in H$:

$$R(h) \leqslant \frac{1}{n}\sum_{i=1}^{n} \ell(m_h(\mathbf{x}_i, y_i)) + 2LK\mathfrak{R}^*_n(\mathcal{F}_{\mathcal{H}}) + 3B_\ell\sqrt{\frac{\log\frac{2}{\delta}}{2n}},$$

where $\mathcal{F}_{\mathcal{H}} = \{f : \mathbf{x} \mapsto h(\mathbf{x}, y) : y \in \mathcal{Y}, h \in \mathcal{H}\}$.

Note that, up-to a constant similar bounds were obtained by Kuznetsov, Mohri, and Syed (2015) and Maximov and Reshetova (2016).

Appendix B. Full Proofs

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 following inequality holds with probability at least $1 - \delta$:*

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

Proof. As the function Δ_n (Eq. 9) is a metric (Appendix, Th. 6); for any labeled training set $S_\ell \subseteq (\mathcal{X} \times \mathcal{Y})^n$ and any cluterings $\mathcal{A}_Z, \mathcal{A}_{Z'}$ found by the algorithm $\mathcal{A}$ over the sets Z, Z', we have by the triangle inequality :

$$\Delta_n(\mathcal{A}_Z, \mathcal{A}^\star, S_\ell) \leqslant \Delta_n(\mathcal{A}_Z, \mathcal{A}_{Z'}, S_\ell) + \Delta_n(\mathcal{A}_{Z'}, \mathcal{A}^\star, S_\ell),$$

hence by the non-negativity of the distance function we have :

$$|\Delta_n(\mathcal{A}_Z, \mathcal{A}^\star, S_\ell) - \Delta_n(\mathcal{A}_{Z'}, \mathcal{A}^\star, S_\ell)| \leqslant \Delta_n(\mathcal{A}_Z, \mathcal{A}_{Z'}, S_\ell). \tag{16}$$

Consider the following multivariate function defined over unlabeled training sets of size u;

$$\begin{aligned} \phi : \mathcal{X}^u &\to \mathbb{R} \\ Z &\mapsto \mathbb{E}_{S_\ell\sim\mathcal{D}^n}[\Delta_n(\mathcal{A}_Z, \mathcal{A}^\star, S_\ell)]. \end{aligned}$$

For any unlabeled training sets, S_u and S'_u drawn i.i.d. with respect to the marginal $\mathcal{D}_\mathcal{X}$ that differ only in one observation we have :

$$\begin{aligned} \left|\phi(S_u) - \phi(S'_u)\right| &= \left|\mathbb{E}_{S_\ell\sim\mathcal{D}^n}\left(\Delta_n(\mathcal{A}_{S_u}, \mathcal{A}^\star, S_\ell) - \Delta_n(\mathcal{A}_{S'_u}, \mathcal{A}^\star, S_\ell)\right)\right| \\ &\leqslant \mathbb{E}_{S_\ell\sim\mathcal{D}^n}\left|\left(\Delta_n(\mathcal{A}_{S_u}, \mathcal{A}^\star, S_\ell) - \Delta_n(\mathcal{A}_{S'_u}, \mathcal{A}^\star, S_\ell)\right)\right| \quad (17) \\ &\leqslant \mathbb{E}_{S_\ell\sim\mathcal{D}^n}\left[\Delta_n(\mathcal{A}_{S_u}, \mathcal{A}_{S'_u}, S_\ell)\right] = \Delta(\mathcal{A}_{S_u}, \mathcal{A}_{S'_u}) \leqslant \frac{L}{u}. \quad (18) \end{aligned}$$

where (Eq. 17) is due to the triangle inequality with absolute value; and (Eq. 18) results from (Eq. 16) and the bounded-difference property of algorithm $\mathcal{A}$ (Eq. 10).

Then by McDiarmid's inequality (Appendix, Th. 5) for any $\epsilon > 0$ we get :

$$\mathbb{P}\left[\phi(S_u) - \mathbb{E}_{S_u \sim \mathcal{D}_{\mathcal{X}}^u} \phi(S_u) \geqslant \epsilon\right] \leqslant \exp\left(-\frac{2\epsilon^2 u}{L^2}\right)$$

Setting the right-hand side to be $\delta/2$, and solving for ϵ, we obtain that with probability at least $1-\frac{\delta}{2}$:

$$\phi(S_u) \leqslant \mathbb{E}_{S_u \sim \mathcal{D}_{\mathcal{X}}^u}[\phi(S_u)] + L\sqrt{\frac{\log \frac{2}{\delta}}{2u}} \leqslant \frac{L}{u} + L\sqrt{\frac{\log \frac{2}{\delta}}{2u}}. \tag{19}$$

Where the last inequality is due to the stability of the clustering algorithm $\mathcal{A}$ (Eq. 11). Furthermore, by bounding $\phi(S_u) = \mathbb{E}_{S_\ell \sim \mathcal{D}^n}[\Delta_n(\mathcal{A}_{S_u}, \mathcal{A}^\star, S_\ell)]$ in terms of S_ℓ using again the McDiarmid inequality we have for any $\epsilon > 0$:

$$\mathbb{P}\left[\Delta_n(\mathcal{A}_{S_u}, \mathcal{A}^\star, S_\ell) - \phi(S_u) \geqslant \epsilon\right] \leqslant e^{-2n\epsilon^2},$$

Indeed, if we consider the multivariate function $\psi : S_\ell \mapsto \Delta_n(\mathcal{A}_{S_u}, \mathcal{A}^\star, S_\ell)$; changing a single labeled observation in S_ℓ could not change $\Delta_n(\mathcal{A}_{S_u}, \mathcal{A}^\star, S_\ell)$ on more than $1/n$ by definition (Eq. 9). Hence, by setting the right-hand side to be $\delta/2$, and solving for ϵ, we obtain that with probability greater than $1 - \frac{\delta}{2}$:

$$\Delta_n(\mathcal{A}_{S_u}, \mathcal{A}^\star, S_\ell) \leqslant \phi(S_u) + \sqrt{\frac{\log \frac{2}{\delta}}{2n}}. \tag{20}$$

Applying the union bound on both inequalities (Eq. 19) and (Eq. 20), we finally get that for any labeled and unlabeled training sets S_ℓ and S_u and with probability at least $1 - \delta$:

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

Lemma 2 *Let $\mathcal{H} \subseteq \mathbb{R}^{\mathcal{X} \times \mathcal{Y}}$ be a hypothesis set where $\mathcal{Y} = \{1, \ldots, K\}$, and let $S_\ell = (\mathbf{x}_i, y_i)_{1 \leqslant i \leqslant n}$ and $S_u = (\mathbf{x}_{n+i})_{1 \leqslant i \leqslant u}$ 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 $\rho > 0$, $\kappa \in \{1, \ldots, K\}$ then for any $1 > \delta > 0$, the following multi-class classification generalization error bound holds with probability at least $1 - \delta$ for all $h \in \mathcal{H}$ learned by algorithm 1 over a single κ-uniformly bounded cluster $\mathcal{C}_j \in \mathcal{C}_\kappa(\eta)$ derived from S_u by a clustering algorithm $\mathcal{A}_{S_u}$ that partitions the input space into G clusters :*

$$\begin{aligned} R(h, \mathcal{C}_j) \leq \widehat{R}_\rho(h, \mathcal{C}_j) &+ \frac{\eta}{G} + \frac{2\kappa}{\rho}\mathfrak{R}^*_{n,j}(\mathcal{F}_\mathcal{H}) + \frac{2K}{\rho}\mathfrak{R}^*_{u,j}(\mathcal{F}_\mathcal{H}) \\ &+ 5\sqrt{\frac{\kappa n_\eta(j) \log \frac{8K}{\delta}}{2n^2}} + 5\sqrt{\frac{\kappa u_\eta(j) \log \frac{8K}{\delta}}{2u^2}} + \frac{7 \log \frac{8}{\delta}}{3(n-1)} + \frac{7 \log \frac{8}{\delta}}{3(u-1)}, \end{aligned}$$

where $n_\eta(j) = |S_\ell \cap \mathcal{C}_j|$, $u_\eta(j) = |S_u \cap \mathcal{C}_j|$, $\mathfrak{R}^*_{n,j} = \mathbb{E}_\sigma \sup\limits_{f \in \mathcal{F}_\mathcal{H}} \frac{2}{n} \sum_{\mathbf{x}_i \in S_\ell \cap \mathcal{C}_j} \sigma_i f(\mathbf{x}_i)$, and $\mathfrak{R}^*_{u,j} = \mathbb{E}_\sigma \sup\limits_{f \in \mathcal{F}_\mathcal{H}} \frac{2}{u} \sum_{\mathbf{x}_i \in S_u \cap \mathcal{C}_j} \sigma_i f(\mathbf{x}_i)$.

Proof. We start with the decomposition of the risk estimated in a single κ-uniformly bounded cluster $\mathcal{C}_j \in \mathcal{C}_\kappa(\eta)$, by considering two situations where the prediction $\mu_h(\mathbf{x}) = \arg\max_{y\in\mathcal{Y}} h(\mathbf{x}, y)$ falls within any set of confident clusters and without them respectively:

$$R(h, \mathcal{C}_j) = \mathbb{E}[\mu_h(\mathbf{x}) \neq y \wedge \mathbf{x} \in \mathcal{C}_j] \leqslant \mathbb{E}_{(\mathbf{x},y)\sim\mathcal{D}}[\mu_h(\mathbf{x}) \neq y \wedge \mu_h(\mathbf{x}) = \mu_h(\mathbf{x}, \mathcal{Y}'_\kappa) \wedge \mathbf{x} \in \mathcal{C}_j] + \\ \mathbb{E}_{(\mathbf{x},y)\sim\mathcal{D}}[\mu_h(\mathbf{x}) \neq y \wedge \mu_h(\mathbf{x}) \neq \mu_h(\mathbf{x}, \mathcal{Y}'_\kappa) \wedge \mathbf{x} \in \mathcal{C}_j] \quad (21)$$

where $\mu_h(\mathbf{x}, \mathcal{Y}'_\kappa) = \arg\max_{y\in\mathcal{Y}'_\kappa} h(\mathbf{x}, y)$ and $\mathcal{Y}'_\kappa \subseteq \mathcal{Y}$, $|\mathcal{Y}'_\kappa| \leqslant \kappa$.

The first term in the inequality above involves the margin of examples and it can be upper-bounded using the definition of the ρ-margin loss (Eq. 8) estimated over the labeled examples that are in cluster $\mathcal{C}_j$:

$$\mathbb{E}_{(\mathbf{x},y)\sim\mathcal{D}}[\mu_h(\mathbf{x}) \neq y \wedge \mu_h(\mathbf{x}) = \mu_h(\mathbf{x}, \mathcal{Y}'_\kappa) \wedge \mathbf{x} \in \mathcal{C}_j] = \mathbb{E}_{(\mathbf{x},y)\sim\mathcal{D}}[\mu_h(\mathbf{x}, \mathcal{Y}'_\kappa) \neq y \wedge \mathbf{x} \in \mathcal{C}_j] \\ \leqslant \mathbb{E}_{S_\ell\sim\mathcal{D}^n}[\Phi_\rho(m_h(\mathbf{x}, y, \mathcal{Y}'_\kappa)) \wedge \mathbf{x} \in \mathcal{C}_j], \quad (22)$$

where $m_h(\mathbf{x}, y, \mathcal{Y}'_\kappa) = h(\mathbf{x}, y) - \max_{y'\in\mathcal{Y}'_\kappa\setminus\{y\}} h(\mathbf{x}, y'), \mathbf{x} \in \mathcal{C}_j$.

Expected risk over a single cluster $\mathcal{C}_j$ can be decomposed through conditional risk as :

$$\mathbb{E}_{S_\ell\sim\mathcal{D}^n}[\Phi_\rho(m_h(\mathbf{x}, y, \mathcal{Y}'_\kappa)) \wedge \mathbf{x} \in \mathcal{C}_j] = \mathbb{E}_{S_\ell\sim\mathcal{D}^n}[\Phi_\rho(m_h(\mathbf{x}, y, \mathcal{Y}'_\kappa)) \big| \mathbf{x} \in \mathcal{C}_j] \times \\ \mathbb{E}_{S_\ell\sim\mathcal{D}^n}[\mathbf{x} \in \mathcal{C}_j] \quad (23)$$

From the data-dependent Bennett's inequality (Appendix A, Thm. 7), we have with probability at least $1 - \delta/4$:

$$\mathbb{E}_{S_\ell\sim\mathcal{D}^n}[\mathbf{x} \in \cap\mathcal{C}_j] \leqslant \frac{n_\eta(j)}{n} + \sqrt{\frac{2n_\eta(j)\log\frac{8}{\delta}}{n^2}} + \frac{7\log\frac{8}{\delta}}{3(n-1)}, \quad (24)$$

where $n_\eta(j) = |S_\ell \cap \mathcal{C}_j|$, and the sample variance, which is upper-bounded by :

$$V_n(\mathbf{x} \in \mathcal{C}_j) = \frac{n_\eta(j)(n - n_\eta(j))}{n(n-1)} \leqslant \frac{n_\eta(j)}{n}.$$

Since $0 \leqslant \Phi_\rho(\cdot) \leqslant 1$ and so $0 \leqslant \mathbb{E}_{(\mathbf{x},y)\sim\mathcal{D}}[\Phi_\rho(m_h(\mathbf{x}, y, \mathcal{Y}'_\kappa)) \big| \mathbf{x} \in S_\ell \cap \mathcal{C}_j] \leqslant 1$, we have from (23) and (24) with probability at least $1 - \delta/4$:

$$\mathbb{E}_{S_\ell\sim\mathcal{D}^n}[\Phi_\rho(m_h(\mathbf{x}, y, \mathcal{Y}'_\kappa) \wedge \mathbf{x} \in \mathcal{C}_j] \leqslant \frac{n_\eta(j)}{n}\mathbb{E}_{S_\ell\sim\mathcal{D}^n}[\Phi_\rho(m_h(\mathbf{x}, y, \mathcal{Y}'_\kappa)) \big| \mathbf{x} \in \mathcal{C}_j] + \\ \sqrt{\frac{2n_\eta(j)\log\frac{8}{\delta}}{n^2}} + \frac{7\log\frac{8}{\delta}}{3(n-1)} \quad (25)$$

Further, the ρ-margin loss function $\Phi_\rho(\cdot)$ (Eq. (8)) is $1/\rho$-Lipschitz, from the multi-class classification generalization bound proposed by Lei et al. (2015) (Appendix A, Thm. 11); it then comes that for any fixed set $\mathcal{Y}'_\kappa \subset \mathcal{Y}$, $|\mathcal{Y}'_\kappa| \leqslant \kappa$ and any $1 > \delta > 0$ with probability at least $1 - \delta/4K^\kappa$ we have for all $h \in \mathcal{H}$:

$$\mathbb{E}_{S_\ell \sim \mathcal{D}^n}[\Phi_\rho(m_h(\mathbf{x}, y, \mathcal{Y}'_\kappa) \big| \mathbf{x} \in \mathcal{C}_j]$$
$$\leqslant \frac{1}{n_\eta(j)} \sum_{(\mathbf{x},y)\in S_\ell \cap \mathcal{C}_j} \Phi_\rho(m_h(\mathbf{x}, y, \mathcal{Y}'_\kappa)) + \frac{2\kappa}{\rho}\mathfrak{R}^*_{n_\eta(j)}(\mathcal{F}) + 3\sqrt{\frac{\log \frac{8K^\kappa}{\delta}}{2n_\eta(j)}}$$
$$\leqslant \frac{1}{n_\eta(j)} \sum_{(\mathbf{x},y)\in S_\ell \cap \mathcal{C}_j} \Phi_\rho(m_h(\mathbf{x}, y, \mathcal{Y}'_\kappa)) + \frac{2\kappa}{\rho}\mathfrak{R}^*_{n_\eta(j)}(\mathcal{F}) + 3\sqrt{\frac{\kappa \log \frac{8K}{\delta}}{2n_\eta(j)}}, \qquad (26)$$

where, $\mathfrak{R}^*_{n_\eta(j)}(\mathcal{F}) = \mathbb{E}_\sigma \sup\limits_{f\in\mathcal{F}_\mathcal{H}} \frac{2}{n_\eta(j)} \sum_{\mathbf{x}_i \in S_\ell \cap \mathcal{C}_j} \sigma_i f(\mathbf{x}_i)$.

Now for any possible set of κ predominant classes $\mathcal{Y}_\kappa$ in $\mathcal{C}_j$, and using the union bound and the inequality $\sum_{i=1}^{k} \binom{K}{i} \leqslant 2K^\kappa$, it comes from (25) and (26) and the union bound, we have with probability at least $1 - \delta/2$:

$$\mathbb{E}_{S_\ell \sim \mathcal{D}^n}[\Phi_\rho(m_h(\mathbf{x}, y, \mathcal{Y}_\kappa)) \wedge \mathbf{x} \in \mathcal{C}_j] \leqslant$$
$$\frac{1}{n} \sum_{(\mathbf{x},y)\in S_\ell \cap \mathcal{C}_j} \Phi_\rho(m_h(\mathbf{x}, y, \mathcal{Y}_\kappa)) + \frac{2\kappa}{\rho}\mathfrak{R}^*_{n,j}(\mathcal{F}) + 5\sqrt{\frac{\kappa n_\eta(j) \log \frac{8K}{\delta}}{2n^2}} + \frac{7 \log \frac{8}{\delta}}{3(n-1)}, \qquad (27)$$

Where $\mathfrak{R}^*_{n,j} = \mathbb{E}_\sigma \sup\limits_{f\in\mathcal{F}_\mathcal{H}} \frac{2}{n} \sum_{\mathbf{x}_i \in S_\ell \cap \mathcal{C}_j} \sigma_i f(\mathbf{x}_i)$. By decomposing the sum in the first term of the above inequality, and considering the two cases where the class label y is within or without $\mathcal{Y}_\kappa$:

$$\sum_{(\mathbf{x},y)\in S_\ell \cap \mathcal{C}_j} \Phi_\rho(m_h(\mathbf{x}, y, \mathcal{Y}_\kappa)) \leqslant \sum_{(\mathbf{x},y)\in S_\ell \cap \mathcal{C}_j \wedge y \in \mathcal{Y}_\kappa} \Phi_\rho(m_h(\mathbf{x}, y, \mathcal{Y}_\kappa)) + \sum_{(\mathbf{x},y)\in S_\ell \cap \mathcal{C}_j \wedge y \notin \mathcal{Y}_\kappa} \Phi_\rho(m_h(\mathbf{x}, y, \mathcal{Y}_\kappa)),$$

Here we are in the case where $\mu_h(\mathbf{x}) = \mu_h(\mathbf{x}, \mathcal{Y}_\kappa)$ (Eq. 21) so, $\forall(\mathbf{x}, y) \in S_\ell \cap \mathcal{C}_j \wedge y \in \mathcal{Y}_\kappa$, $\Phi_\rho(m_h(\mathbf{x}, y, \mathcal{Y}_\kappa)) = \Phi_\rho(m_h(\mathbf{x}, y))$, and $\forall(\mathbf{x}, y) \in S_\ell \cap \mathcal{C}_j \wedge y \notin \mathcal{Y}_\kappa, \Phi_\rho(m_h(\mathbf{x}, y, \mathcal{Y}_\kappa)) \leqslant \mathbb{1}_{y\notin\mathcal{Y}_\kappa \wedge \mathbf{x}\in\mathcal{C}_j}$. Hence, for any sample S_ℓ and a set of predominant classes $\mathcal{Y}_\kappa$ we have

$$\frac{1}{n} \sum_{(\mathbf{x},y)\in S_\ell \cap \mathcal{C}_j} \Phi_\rho(m_h(\mathbf{x}, y, \mathcal{Y}_\kappa)) \leqslant \frac{1}{n} \sum_{(\mathbf{x},y)\in S_\ell \cap \mathcal{C}_j} \Phi_\rho(m_h(\mathbf{x}, y)) + \frac{1}{n} \sum_{(\mathbf{x},y)\in S_\ell} \mathbb{1}_{y\notin\mathcal{Y}_\kappa \wedge \mathbf{x}\in\mathcal{C}_j}$$
$$\leqslant \frac{1}{n} \sum_{(\mathbf{x},y)\in S_\ell \cap \mathcal{C}_j} \Phi_\rho(m_h(\mathbf{x}, y)) + \frac{1}{n} \sum_{(\mathbf{x},y)\in S_\ell} \mathbb{1}_{y\notin\mathcal{Y}_\kappa \wedge \mathbf{x}\in\mathcal{C}_j}.$$

From definition (1) we have $\frac{1}{n}\sum_{(\mathbf{x},y)\in S_\ell} \mathbb{1}_{y\notin\mathcal{Y}_\kappa \wedge \mathbf{x}\in\mathcal{C}_j} \leqslant \eta/G$, and so

$$\mathbb{E}_{(\mathbf{x},y)\sim\mathcal{D}}[\mu_h(\mathbf{x}) \neq y \wedge \mu_h(\mathbf{x}) = \mu_h(\mathbf{x}, \mathcal{Y}_\kappa) \wedge \mathbf{x} \in \mathcal{C}_j] \leqslant$$
$$\frac{1}{n} \sum_{(\mathbf{x},y)\in S_\ell \cap \mathcal{C}_j} \Phi_\rho(m_h(\mathbf{x}, y)) + \frac{\eta}{G} + \frac{2\kappa}{\rho}\mathfrak{R}^*_{n,j}(\mathcal{F}) + 5\sqrt{\frac{\kappa n_\eta(j) \log \frac{8K}{\delta}}{2n^2}} + \frac{7 \log \frac{8}{\delta}}{3(n-1)}. \qquad (28)$$

Further, the second term in inequality (21) for any set $\mathcal{Y}_\kappa \subset \mathcal{Y}$, $|\mathcal{Y}_\kappa| \leqslant \kappa$ can be upperbounded using unlabeled data that are in cluster $\mathcal{C}_j$:

$$\begin{aligned}\mathbb{E}[\mu_h(\mathbf{x}) \neq y \wedge \mu_h(\mathbf{x}) \neq \mu_h(\mathbf{x}, \mathcal{Y}_\kappa) \wedge \mathbf{x} \in \mathcal{C}_j] &\leqslant \mathbb{E}_{S_u \sim \mathcal{D}_{\mathcal{X}}^u}[\mu_h(\mathbf{x}) \neq \mu_h(\mathbf{x}, \mathcal{Y}_\kappa) \wedge \mathbf{x} \in \mathcal{C}_j] \\ &\leqslant \mathbb{E}_{S_u \sim \mathcal{D}_{\mathcal{X}}^u}[\Phi_\rho(m'_h(\mathbf{x}, \mathcal{Y}_\kappa)) \wedge \mathbf{x} \in \mathcal{C}_j],\end{aligned}$$

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

As the ρ-margin loss has its values in $[0, 1]$, from the standard Rademacher complexity bound (Appendix A, Thm. 9) over i.i.d. sample $S_u \cap \mathcal{C}_j$, for any $0 > \delta > 1$ and $\mathcal{Y}_\kappa \subseteq \mathcal{Y}$ it comes that with probability at least $1 - \delta/4$:

$$\begin{aligned}\mathbb{E}_{S_u \sim \mathcal{D}_{\mathcal{X}}^u}[\Phi_\rho(m'_h(\mathbf{x}, \mathcal{Y}_\kappa)) \big| \mathbf{x} \in \mathcal{C}_j] \leqslant & \frac{1}{u_\eta(j)} \sum_{\mathbf{x} \in \mathcal{C}_j \cap S_u} \Phi_\rho(m'_h(\mathbf{x}, \mathcal{Y}_\kappa)) + \\ & \sum_{\mathcal{C}_j \in \mathcal{C}_\kappa(\eta)} \mathbb{E}_\sigma \sup_{f \in \mathcal{G}_1^{\mathcal{C}_j} \cup \mathcal{G}_2^{\mathcal{C}_j}} \frac{2}{u_\eta(j)} \sum_{\mathbf{x}_i \in S_u \cap \mathcal{C}_j} \sigma_i f(\mathbf{x}_i) + 3\sqrt{\frac{\log \frac{8K^\kappa}{\delta}}{2u_\eta(j)}},\end{aligned} \quad (29)$$

where $\mathcal{G}_1^{\mathcal{C}_j} = \{\max_{y \in \mathcal{Y}_\kappa(\mathcal{C}_j)} h(\mathbf{x}, y), h \in \mathcal{F}_\mathcal{H}\}$ and $\mathcal{G}_2^{\mathcal{C}_j} = \{\max_{y \notin \mathcal{Y}_\kappa(\mathcal{C}_j)} h(\mathbf{x}, y), h \in \mathcal{F}_\mathcal{H}\}$. Due to the monotonicity of supremum, we have for any $\mathcal{C}_j \in \mathcal{C}_\kappa(\eta)$:

$$\begin{aligned}&\mathbb{E}_\sigma \sup_{f \in \mathcal{G}_1^{\mathcal{C}_j} \cup \mathcal{G}_2^{\mathcal{C}_j}} \frac{2}{u_\eta(j)} \sum_{\mathbf{x}_i \in S_u \cap \mathcal{C}_j} \sigma_i f(\mathbf{x}_i) \\ &\qquad \leqslant \mathbb{E}_\sigma \sup_{f \in \mathcal{G}_1^{\mathcal{C}_j}} \frac{2}{u_\eta(j)} \sum_{\mathbf{x}_i \in S_u \cap \mathcal{C}_j} \sigma_i f(\mathbf{x}_i) + \mathbb{E}_\sigma \sup_{f \in \mathcal{G}_2^{\mathcal{C}_j}} \frac{2}{u_\eta(j)} \sum_{\mathbf{x}_i \in S_u \cap \mathcal{C}_j} \sigma_i f(\mathbf{x}_i)\end{aligned}$$

By Lemma 8 (Appendix A) we have :

$$\begin{aligned}&\mathbb{E}_\sigma \sup_{f \in \mathcal{G}_1^{\mathcal{C}_j}} \frac{2}{u_\eta(j)} \sum_{\mathbf{x}_i \in S_u \cap \mathcal{C}_j} \sigma_i f(\mathbf{x}_i) + \mathbb{E}_\sigma \sup_{f \in \mathcal{G}_2^{\mathcal{C}_j}} \frac{2}{u_\eta(j)} \sum_{\mathbf{x}_i \in S_u \cap \mathcal{C}_j} \sigma_i f(\mathbf{x}_i) \\ &\qquad \leqslant K \mathbb{E}_\sigma \sup_{f \in \mathcal{F}} \frac{2}{u_\eta(j)} \sum_{\mathbf{x}_i \in S_u \cap \mathcal{C}_j} \sigma_i f(\mathbf{x}_i)\end{aligned}$$

Hence,

$$\mathbb{E}_{S_u \sim \mathcal{D}_{\mathcal{X}}^u}[\Phi_\rho(m'_h(\mathbf{x}, \mathcal{Y}_\kappa)) \big| \mathbf{x} \in \mathcal{C}_j] \leqslant \frac{1}{u_\eta(j)} \sum_{\mathbf{x} \in \mathcal{C}_j} \Phi_\rho(m'_h(\mathbf{x}, \mathcal{Y}_\kappa)) + \frac{2K}{\rho} \mathfrak{R}^*_{u_\eta(j)}(\mathcal{F}) + 3\sqrt{\frac{\kappa \log \frac{8K}{\delta}}{2u_\eta(j)}}, \quad (30)$$

where $\mathfrak{R}^*_{u_\eta(j)}(\mathcal{F}) = \mathbb{E}_\sigma \sup_{f \in \mathcal{F}} \frac{2}{u_\eta(j)} \sum_{\mathbf{x}_i \in S_u \cap \mathcal{C}_j} \sigma_i f(\mathbf{x}_i)$. Similarly to (25) we have with probability at least $1 - \delta/4$:

$$\begin{aligned}\mathbb{E}_{S_u \sim \mathcal{D}_{\mathcal{X}}^u}[\Phi_\rho(m'_h(\mathbf{x}, y, \mathcal{Y}'_\kappa) \wedge \mathbf{x} \in \mathcal{C}_j] \leqslant & \frac{u_\eta(j)}{u} \mathbb{E}_{S_u \sim \mathcal{D}_{\mathcal{X}}^u}[\Phi_\rho(m'_h(\mathbf{x}, y, \mathcal{Y}'_\kappa)) \big| \mathbf{x} \in \mathcal{C}_j] + \\ & \sqrt{\frac{2u_\eta(j) \log \frac{8}{\delta}}{u^2}} + \frac{7 \log \frac{8}{\delta}}{3(u-1)}\end{aligned} \quad (31)$$

Thus, by (30) and (31), and the union bound we have with probability at least $1-\delta/2$:

$$\mathbb{E}_{\mathbf{x}\sim\mathcal{D}_{\mathcal{X}}}[\Phi_\rho(m'_h(\mathbf{x},\mathcal{Y}_\kappa))\wedge \mathbf{x}\in\mathcal{C}_j] \leqslant \frac{1}{u}\sum_{\mathbf{x}\in\mathcal{C}_j}\Phi_\rho(m'_h(\mathbf{x},\mathcal{Y}_\kappa)) + \frac{2K}{\rho}\mathfrak{R}^*_{u,j}(\mathcal{F})+ 5\sqrt{\frac{\kappa u_\eta(j)\log\frac{8K}{\delta}}{2u^2}} + \frac{7\log\frac{8}{\delta}}{3(u-1)} \tag{32}$$

The statement of the Lemma follows from the inequalities (21), (28), (32), and the union bound. □

Theorem 3 *Let $\mathcal{H}\subseteq\mathbb{R}^{\mathcal{X}\times\mathcal{Y}}$ be a hypothesis set where $\mathcal{Y}=\{1,\dots,K\}$, 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 $\kappa\in\{1,\dots,K\}$, 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 $\eta\le\Delta_n(\mathcal{A}_{S_u},\mathcal{A}^\star,S_\ell)$, 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 $1-\delta$:*

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

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

Proof. Let $\Pi_{S_u}=\{\mathcal{C}_1,\dots,\mathcal{C}_G\}$ be a set of disjoint clusters found by $\mathcal{A}_{S_u}$. We decompose the risk of a classifier by considering the two exclusive cases whether the misclassification error occurs inside or outside the set of η-confident clusters :

$$R(h)=\mathbb{E}_{(\mathbf{x},y)\sim\mathcal{D}}[\mu_h(\mathbf{x})\ne y]=\sum_{\mathcal{C}_j\in\mathcal{C}_\kappa(\eta)}\mathbb{E}_{(\mathbf{x},y)\sim\mathcal{D}}[\mu_h(\mathbf{x})\ne y\wedge\mathbf{x}\in\mathcal{C}_j]+\sum_{\mathcal{C}_j\notin\mathcal{C}_\kappa(\eta)}\mathbb{E}_{(\mathbf{x},y)\sim\mathcal{D}}[\mu_h(\mathbf{x})\ne y\wedge\mathbf{x}\in\mathcal{C}_j]. \tag{33}$$

First, we bound the risk over the set of confident clusters. For any cluster $\mathcal{C}_j$ in $\mathcal{C}_\kappa(\eta)$ and any set of confident clusters $\mathcal{Y}_\kappa(\mathcal{C}_j)$ within it, from lemma 2 we have with probability at least $1-\frac{4\delta}{7G}$:

$$R(h,\mathcal{C}_j)=\mathbb{E}_{(\mathbf{x},y)\sim\mathcal{D}}[\mu_h(\mathbf{x})\ne y\wedge\mathbf{x}\in\mathcal{C}_j]\leqslant\widehat{R}_\rho(h,\mathcal{C}_j)+\frac{\eta}{G}+\frac{2\kappa}{\rho}\mathfrak{R}^*_{n,j}(\mathcal{F}_\mathcal{H})+\frac{2K}{\rho}\mathfrak{R}^*_{u,j}(\mathcal{F}_\mathcal{H}) +5\sqrt{\frac{\kappa n_\eta(j)\log\frac{14G}{\delta}}{2n^2}}+5\sqrt{\frac{\kappa u_\eta(j)\log\frac{14G}{\delta}}{2u^2}}+\frac{7\log\frac{14G}{\delta}}{3(n-1)}+\frac{7\log\frac{14G}{\delta}}{3(u-1)},$$

where $n_\eta(j) = |S_\ell \cap \mathcal{C}_j|$, and $\mathfrak{R}^*_{n,j}(\mathcal{F}) = \mathbb{E}_{\sigma,S_\ell} \sup_{f\in\mathcal{F}_\mathcal{H}} \frac{2}{n}\left|\sum_{\mathbf{x}_i\in S_\ell\cap\mathcal{C}_j} \sigma_i f(\mathbf{x}_i)\right|$, and $u_\eta(j) = |S_u \cap \mathcal{C}_j|$, and $\mathfrak{R}^*_{u,j}(\mathcal{F}) = \mathbb{E}_{\sigma,S_u} \sup_{f\in\mathcal{F}_\mathcal{H}} \frac{2}{u}\left|\sum_{\mathbf{x}_i\in S_u\cap\mathcal{C}_j} \sigma_i f(\mathbf{x}_i)\right|$. Summing up over all clusters it comes

$$\sum_{\mathcal{C}_j\in\mathcal{C}_\kappa(\eta)} R(h,\mathcal{C}_j) \leq \sum_{\mathcal{C}_j\in\mathcal{C}_\kappa(\eta)} \widehat{R}_\rho(h,\mathcal{C}_j) + \eta + \frac{2\kappa}{\rho}\mathfrak{R}^*_n(\mathcal{F}_\mathcal{H}) + \frac{2K}{\rho}\mathfrak{R}^*_u(\mathcal{F}_\mathcal{H}) + 5\sum_{j=1}^{G}\sqrt{\frac{\kappa n_\eta(j)\log\frac{14G}{\delta}}{2n^2}} + 5\sum_{j=1}^{G}\sqrt{\frac{\kappa u_\eta(j)\log\frac{14G}{\delta}}{2u^2}} + \frac{7G\log\frac{14G}{\delta}}{3(n-1)} + \frac{7G\log\frac{14G}{\delta}}{3(u-1)}.$$

By the Cauchy–Schwarz inequality $(\sum_{i=1}^G a_i b_i)^2 \leqslant (\sum_{i=1}^G a_i^2)(\sum_{i=1}^G b_i^2)$, then by fixing $b_i = 1, \forall i \in \{1,\dots,G\}$, we can bound the two last terms of the right hand side inequality, and get

$$\sum_{\mathcal{C}_j\in\mathcal{C}_\kappa(\eta)} R(h,\mathcal{C}_j) \leq \sum_{\mathcal{C}_j\in\mathcal{C}_\kappa(\eta)} \widehat{R}_\rho(h,\mathcal{C}_j) + \eta + \frac{2\kappa}{\rho}\mathfrak{R}^*_n(\mathcal{F}_\mathcal{H}) + \frac{2K}{\rho}\mathfrak{R}^*_u(\mathcal{F}_\mathcal{H}) + 5\sqrt{\frac{Gn_\eta\kappa\log\frac{14KG}{\delta}}{2n^2}} + 5\sqrt{\frac{Gu_\eta\kappa\log\frac{14KG}{\delta}}{2u^2}} + \frac{7G\log\frac{14G}{\delta}}{3(n-1)} + \frac{7G\log\frac{14G}{\delta}}{3(u-1)}, \tag{34}$$

with $n^*_\eta = \sum_{\mathcal{C}_j\in\mathcal{C}_\kappa(\eta)} = n^*_\eta(j)$ and $u^*_\eta = \sum_{\mathcal{C}_j\in\mathcal{C}_\kappa(\eta)} = u^*_\eta(j)$.

From the inequality $\eta \leq \Delta_n(\mathcal{A}_{S_u}, \mathcal{A}^\star, S_\ell)$ and Lemma 1, the following upper-bound holds with probability at least $1 - \frac{\delta}{7}$:

$$\eta \leq \frac{L}{u} + L\sqrt{\frac{\log\frac{14}{\delta}}{2u}} + \sqrt{\frac{\log\frac{14}{\delta}}{2n}}$$

By the inequality $\forall a > 0, b > 0; (a+b)^2 \leqslant 2(a^2+b^2)$ it then comes :

$$\eta \leq \frac{L}{u} + \sqrt{\left(\frac{L^2}{u} + \frac{1}{n}\right)\log\frac{14}{\delta}} \tag{35}$$

Further the risk of classification outside the set of confident clusters can be decomposed as :

$$\mathbb{E}_{(\mathbf{x},y)\sim\mathcal{D}}[\mu_h(\mathbf{x}) \neq y \wedge \mathbf{x} \in S_\ell \setminus \mathcal{C}_\kappa(\eta)] = \mathbb{E}_{(\mathbf{x},y)\sim\mathcal{D}}[\mu_h(\mathbf{x}) \neq y \mid \mathbf{x} \in S_\ell \setminus \mathcal{C}_\kappa(\eta)] \times \mathbb{E}_{(\mathbf{x},y)\sim\mathcal{D}}[\mathbf{x} \in S_\ell \setminus \mathcal{C}_\kappa(\eta)]. \tag{36}$$

Similarly to the previous development, and from the multi-class classification generalization bound and the Data-dependent Bennett's inequality (Appendix A, Thm. 11 & 7), the above risk is upper-bounded with probability at least $1 - 2\frac{\delta}{7}$ by :

$$\mathbb{E}_{(\mathbf{x},y)\sim\mathcal{D}}[\mu_h(\mathbf{x}) \neq y \mid \mathbf{x} \in S_\ell \setminus \mathcal{C}_\kappa(\eta)] \leqslant \frac{1}{n}\sum_{(\mathbf{x},y)\in S_\ell\setminus\mathcal{C}_\kappa(\eta)} \Phi_\rho(m_h(\mathbf{x},y)) + \frac{2K}{\rho}\mathfrak{R}_n(\mathcal{F}_\mathcal{H}) + 5\sqrt{\frac{K(n-n_\eta)\log\frac{14K}{\delta}}{2n^2}} + \frac{7\log\frac{14}{\delta}}{3(n-1)} \tag{37}$$

The result then follows from the inequalities, $\forall a > 0, b > 0, c > 0; (a+b+c)^2 \leqslant 3(a^2+b^2+c^2)$; $5\sqrt{3} < 9$; (33), (34), (35), (37) and the union-bound. $\square$

References

Amini, M., Laviolette, F., & Usunier, N. (2008a). A transductive bound for the voted classifier with an application to semi-supervised learning. In *Advances in Neural Information Processing Systems (NIPS 22)*, pp. 65–72.

Amini, M., Truong, T., & Goutte, C. (2008b). A boosting algorithm for learning bipartite ranking functions with partially labeled data. In *Proceedings of the 31st Annual International ACM SIGIR Conference*, pp. 99–106.

Amini, M., & Usunier, N. (2015). *Learning with Partially Labeled and Interdependent Data*. Springer.

Balcan, M., & Blum, A. (2010). A discriminative model for semi-supervised learning. *J. ACM*, *57*(3).

Ben-David, S., Lu, T., & Pál, D. (2008). Does unlabeled data provably help? worst-case analysis of the sample complexity of semi-supervised learning. In *21st Annual Conference on Learning Theory - COLT 2008, Helsinki, Finland, July 9-12, 2008*, pp. 33–44.

Blum, A., & Mitchell, T. M. (1998). Combining labeled and unlabeled sata with co-training. In *Proceedings of the Eleventh Annual Conference on Computational Learning Theory, COLT 1998, Madison, Wisconsin, USA, July 24-26, 1998.*, pp. 92–100.

Bubeck, S., & Luxburg, U. V. (2009). Nearest neighbor clustering: A baseline method for consistent clustering with arbitrary objective functions. *Journal of Machine Learning Research*, *10*, 657–698.

Castelli, V., & Cover, T. M. (1995). On the exponential value of labeled samples. *Pattern Recognition Letters*, *16*(1), 105–111.

Chang, C.-C., & Lin, C.-J. (2011). LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, *2*, 27:1–27:27. Software available at `http://www.csie.ntu.edu.tw/~cjlin/libsvm`.

Chapelle, O., Schölkopf, B., & Zien, A. (2006). *Semi-supervised learning*. MIT press.

El-Yaniv, R., & Pechyony, D. (2009). Transductive rademacher complexity and its applications. *Journal of Artificial Intelligence Research (JAIR)*, *35*, 193–234.

Harchaoui, Z., Douze, M., Paulin, M., Dudík, M., & Malick, J. (2012). Large-scale image classification with trace-norm regularization. In *2012 IEEE Conference on Computer Vision and Pattern Recognition, Providence, RI, USA, June 16-21, 2012*, pp. 3386–3393.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning: data mining, inference and prediction*. Springer.

Joachims, T. (1999). Transductive inference for text classification using support vector machines. In *Proceedings of ICML-99, 16th International Conference on Machine Learning*, pp. 200–209.

Kääriäinen, M. (2005). Generalization error bounds using unlabeled data. In *Learning Theory, 18th Annual Conference on Learning Theory, COLT 2005, Bertinoro, Italy, June 27-30, 2005, Proceedings*, pp. 127–142.

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

Koltchinskii, V., & Panchenko, D. (2002). Empirical margin distributions and bounding the generalization error of combined classifiers. *Annals of Statistics*, 1–50.

Kuznetsov, V., Mohri, M., & Syed, U. (2015). Rademacher complexity margin bounds for learning with a large number of classes. In *ICML Workshop on Extreme Classification: Learning with a Very Large Number of Labels*.

LeCun, Y., Bottou, L., Bengio, Y., & Haffner, P. (2001). Gradient-based learning applied to document recognition. In Haykin, S., & Kosko, B. (Eds.), *Intelligent Signal Processing*, pp. 306–351. IEEE Press.

Lehmann, E. (1975). *Nonparametric Statistical Methods Based on Ranks.* McGraw-Hill, New York.

Lei, Y., Dogan, U., Binder, A., & Kloft, M. (2015). Multi-class svms: From tighter data-dependent generalization bounds to novel algorithms. In *Advances in Neural Information Processing Systems*, pp. 2035–2043.

Leskes, B. (2005). The value of agreement, a new boosting algorithm. In *Learning Theory, 18th Annual Conference on Learning Theory, COLT 2005, Bertinoro, Italy, June 27-30, 2005, Proceedings*, pp. 95–110.

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

Luxburg, U. V., Bousquet, O., & Belkin, M. (2004). On the convergence of spectral clustering on random samples: The normalized case. In *Learning Theory, 17th Annual Conference on Learning Theory, COLT 2004, Banff, Canada, July 1-4, 2004, Proceedings*, pp. 457–471.

Maji, S., & Malik, J. (2009). Fast and accurate digit classification. Tech. rep. UCB/EECS-2009-159, EECS Department, University of California, Berkeley.

Maurer, A., & Pontil, M. (2009). Empirical bernstein bounds and sample-variance penalization. In *COLT 2009 - The 22nd Conference on Learning Theory, Montreal, Quebec, Canada, June 18-21, 2009*.

Maximov, Y., & Reshetova, D. (2016). Tight risk bounds for multi-class margin classifiers. *Pattern Recognition and Image Analysis*, *26*(4), 673–680.

McDiarmid, C. (1989). On the method of bounded differences. In *Surveys in Combinatorics*, No. 141, pp. 148–188.

Mohri, M., Rostamizadeh, A., & Talwalkar, A. (2012). *Foundations of Machine Learning*. Adaptive computation and machine learning. MIT Press.

Rakhlin, A., & Caponnetto, A. (2006). Stability of k-means clustering. In *Advances in Neural Information Processing Systems 19, Proceedings of the Twentieth Annual Conference on Neural Information Processing Systems, Vancouver, British Columbia, Canada, December 4-7, 2006*, pp. 1121–1128.

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

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A. C., & Fei-Fei, L. (2015). ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)*, *115*(3), 211–252.

Seeger, M. (2000). Learning with labeled and unlabeled data. Tech. rep., Institute for Adaptive and Neural Computation, University of Edinburgh.

Shamir, O., & Tishby, N. (2007). Cluster stability for finite samples. In *Advances in Neural Information Processing Systems 20, Proceedings of the Twenty-First Annual Conference on Neural Information Processing Systems, Vancouver, British Columbia, Canada, December 3-6, 2007*, pp. 1297–1304.

Thiagarajan, J. J., Ramamurthy, K. N., & Spanias, A. (2011). Optimality and stability of the k-hyperline clustering algorithm. *Pattern Recognition Letters*, *32*(9), 1299–1304.

Tolstikhin, I. O., Zhivotovskiy, N., & Blanchard, G. (2015). Permutational rademacher complexity - A new complexity measure for transductive learning. In *Algorithmic Learning Theory - 26th International Conference, ALT*, pp. 209–223.

Urner, R., Shalev-Shwartz, S., & Ben-David, S. (2011). Access to unlabeled data can speed up prediction time. In *Proceedings of the 28th International Conference on Machine Learning, ICML 2011, Bellevue, Washington, USA, June 28 - July 2, 2011*, pp. 641–648.

Vapnik, V. (2000). *The nature of statistical learning theory*. Springer Science & Business Media.

Wang, B., Tu, Z., & Tsotsos, J. K. (2013). Dynamic label propagation for semi-supervised multi-class multi-label classification. In *IEEE International Conference on Computer Vision, ICCV 2013, Sydney, Australia, December 1-8, 2013*, pp. 425–432.

Zhou, D., Bousquet, O., Lal, T. N., Weston, J., & Schölkopf, B. (2003). Learning with local and global consistency. In *Advances in Neural Information Processing Systems 16 [Neural Information Processing Systems, NIPS 2003, December 8-13, 2003, Vancouver and Whistler, British Columbia, Canada]*, pp. 321–328.

Zhu, X. (2002). Learning from labeled and unlabeled data with label propagation. Tech. rep. CMU-CALD-02-107, Carnegie Mellon University.

Zhu, X. (2005). Semi-supervised learning literature survey. technical report 1530. Tech. rep., Department of Computer Sciences, University of Wisconsin.