Neural Networks 82 (2016) 62–75

Contents lists available at ScienceDirect

Neural Networks

journal homepage: www.elsevier.com/locate/neunet

A local Vapnik–Chervonenkis complexity

Luca Oneto [a,*], Davide Anguita [a], Sandro Ridella [b]

[a] DIBRIS - University of Genoa, Via Opera Pia 13, I-16145 Genoa, Italy
[b] DITEN - University of Genoa, Via Opera Pia 11A, I-16145 Genoa, Italy

ARTICLE INFO

Article history:
Received 26 January 2016
Received in revised form 19 May 2016
Accepted 1 July 2016
Available online 18 July 2016

Keywords:
Local Rademacher Complexity
Local Vapnik–Chervonenkis entropy
Generalization error bounds
Statistical Learning Theory
Complexity measures

ABSTRACT

We define in this work a new localized version of a Vapnik–Chervonenkis (VC) complexity, namely the Local VC-Entropy, and, building on this new complexity, we derive a new generalization bound for binary classifiers. The Local VC-Entropy-based bound improves on the original Vapnik's results because it is able to discard those functions that, most likely, will not be selected during the learning phase. The result is achieved by applying the localization principle to the original global complexity measure, in the same spirit of the Local Rademacher Complexity. By exploiting and improving a recently developed geometrical framework, we show that it is also possible to relate the Local VC-Entropy to the Local Rademacher Complexity by finding an admissible range for one given the other. In addition, the Local VC-Entropy allows one to reduce the computational requirements that arise when dealing with the Local Rademacher Complexity in binary classification problems.

1. Introduction

In learning systems, the development of effective measures for assessing the complexity of hypothesis classes is fundamental for enabling a precise control of the outcome of the learning process. One of the first attempts was made several decades ago, with the theory developed by V.N. Vapnik and A.Y. Chervonenkis, who proposed, among the others, the well-known VC-Dimension (Vapnik, 1998; Zhang, Bian, Tao, & Lin, 2012; Zhang & Tao, 2013). The VC-Dimension defines the complexity of a hypothesis class as the cardinality of the largest set of points that can be shattered by functions of the class. Unfortunately, the VC-Dimension, like other measures, is a global one, because it takes into account all the functions in the hypothesis class, and, furthermore, is data-independent, because it does not take into account the actual distribution of the data available for learning. As a consequence of targeting this worst-case learning scenario, the VC-Dimension leads to very pessimistic generalization bounds.

In order to deal with the second issue, effective data-dependent complexity measures have been developed, which allow to take into account the actual distribution of the data and produce tighter estimates of the complexity of the class. As an example, data-dependent versions of the VC Complexities have been developed in Boucheron, Lugosi, and Massart (2000), Shawe-Taylor, Bartlett, Williamson, and Anthony (1998) and together with the Rademacher Complexity (Bartlett & Mendelson, 2003; Koltchinskii, 2001) they represent the state-of-the-art tools in this field.

In recent years, the Rademacher Complexity has been further improved, as researchers have succeeded in developing local data-dependent complexity measures (Bartlett, Bousquet, & Mendelson, 2002, 2005; Cortes, Kloft, & Mohri, 2013; Koltchinskii, 2006; Oneto, Ghio, Ridella, & Anguita, 2015b; van de Geer, 2006). Local measures improve over global ones thanks to their ability of taking into account only those functions of the hypothesis class that will be most likely chosen by the learning procedure, i.e. the models with small error. In particular, the Local Rademacher Complexity has shown to be able to accurately capture the nature of the learning process, both from a theoretical point of view (Bartlett et al., 2005; Koltchinskii, 2006; Oneto et al., 2015b) and in real-world applications (Cortes et al., 2013; Kloft & Blanchard, 2011; Lei, Binder, Dogan, & Kloft, 2015; Steinwart & Scovel, 2005).

We propose in this work a localized version of a VC Complexity, namely the Local VC-Entropy, and show how it can be related to the Local Rademacher Complexity through an extension of the geometrical framework presented in Anguita, Ghio, Oneto, and Ridella (2014). Getting more insights on the mechanisms underlying different notions of complexity, and the non-trivial relationships among them, is crucial, both from a theoretical and a practical point of view. In fact, for this reason, the literature

* Corresponding author.
E-mail addresses: Luca.Oneto@unige.it (L. Oneto), Davide.Anguita@unige.it (D. Anguita), Sandro.Ridella@unige.it (S. Ridella).

http://dx.doi.org/10.1016/j.neunet.2016.07.002
0893-6080/

targeting the connections between different complexity measures is quite large (Bousquet, 2002; Ledoux & Talagrand, 1991; Lei, Ding, & Bi, 2015; Massart, 2000; Sauer, 1972; Shelah, 1972; Srebro, Sridharan, & Tewari, 2010; Vapnik, 1998). Finally, we show how to exploit the Local VC-Entropy for bypassing the computational difficulties that arise when computing the Local Rademacher Complexity in binary classification problems.

The localization of the VC-Entropy allows us to introduce the same improvements achieved by the localization of the Rademacher Complexity into the VC Theory as well, like, for example, the derivation of refined generalization bounds with respect to their global counterparts. In fact, based on this new localized notion of complexity, we propose a new generalization bound that does not take into account all the functions in the set but only the ones with small error.

The paper is structured as follows. In Section 2 we introduce the theoretical framework and the two localized notions of complexity, i.e. the Local Rademacher Complexity and the new Local VC-Entropy. In Section 3 we propose a new bound on the generalization error based on the Local VC-Entropy and we show that this new bound is actually able to discard those functions that will be never chosen by the algorithm for classification purposes but are usually considered for estimating the generalization error. Section 4 is devoted to the connections between the Local Rademacher Complexity and the new Local VC-Entropy, by exploiting a geometrical framework introduced in Anguita et al. (2014), which deals only with the global version of these complexities. In Section 5 we show the computational advantages of the Local VC-Entropy with respect to the Local Rademacher Complexity. Section 6 concludes the paper.

2. A local Vapnik–Chervonenkis complexity

Let μ be a probability distribution over $\mathcal{X} \times \mathcal{Y}$ where $\mathcal{Y} = \{\pm 1\}$. We denote as $\mathcal{F}$ a class of $\{\pm 1\}$-valued functions $f \in \mathcal{F}$ on $\mathcal{X}$, and suppose that $\mathcal{D}_n = \{(X_1, Y_1), \ldots, (X_n, Y_n)\}$ with $n > 1$ is sampled according to μ. The accuracy of an $f \in \mathcal{F}$ in representing μ is measured according to the indicator function $I(f(X), Y) = \frac{1-Yf(X)}{2}$, namely $I(f(X), Y) = 0$ if $f(X) = Y$ and $I(f(X), Y) = 1$ if $f(X) \neq Y$. Consequently, the empirical error $\widehat{L}_n(f)$ and the generalization error $L(f)$ of an $f \in \mathcal{F}$ can be defined as:

$$\widehat{L}_n(f) = \frac{1}{n}\sum_{i=1}^{n} I(f(X_i), Y_i), \tag{1}$$

$$L(f) = \mathbb{E}_{(X,Y)} I(f(X), Y). \tag{2}$$

Let us define the following quantity:

$$\mathcal{F}_{\mathcal{D}_n} = \{\{f(X_1), \ldots, f(X_n)\} : f \in \mathcal{F}\}, \tag{3}$$

which is the set of functions restricted to the sample. In other words, $\mathcal{F}_{\mathcal{D}_n}$ is the set of distinct functions distinguishable within $\mathcal{F}$ with respect to the dataset $\mathcal{D}_n$. The VC-Entropy $H_n(\mathcal{F})$ and the Annealed VC-Entropy $A_n(\mathcal{F})$, together with their empirical counterparts $\widehat{H}_n(\mathcal{F})$ and $\widehat{A}_n(\mathcal{F})$ (Vapnik, 1998), are defined as:

$$H_n(\mathcal{F}) = \mathbb{E}_{X_1,\ldots,X_n}\widehat{H}_n(\mathcal{F}), \qquad \widehat{H}_n(\mathcal{F}) = \ln\left(\left|\mathcal{F}_{\mathcal{D}_n}\right|\right), \tag{4}$$

$$A_n(\mathcal{F}) = \ln\left(\mathbb{E}_{X_1,\ldots,X_n}\left|\mathcal{F}_{\mathcal{D}_n}\right|\right), \qquad \widehat{A}_n(\mathcal{F}) = \widehat{H}_n(\mathcal{F}). \tag{5}$$

Let $\{\sigma_1, \ldots, \sigma_n\}$ be n independent Rademacher random variables for which $\mathbb{P}\{\sigma_i = +1\} = \mathbb{P}\{\sigma_i = -1\} = 1/2$. Then, the Rademacher Complexity $\widehat{R}_n(\mathcal{F})$ (Bartlett & Mendelson, 2003; Koltchinskii, 2001), and its deterministic counterpart $R_n(\mathcal{F})$, are defined as:

$$\widehat{R}_n(\mathcal{F}) = \mathbb{E}_{\boldsymbol{\sigma}} \sup_{f\in\mathcal{F}} \frac{2}{n}\sum_{i=1}^{n} I(f(X_i), \sigma_i), \tag{6}$$

$$R_n(\mathcal{F}) = \mathbb{E}_{X_1,\ldots,X_n}\widehat{R}_n(\mathcal{F}). \tag{7}$$

Note that the definition of Rademacher Complexity adopted in this paper is in agreement with the one that appeared in the recent literature (Anguita et al., 2014; Bartlett & Mendelson, 2003; Koltchinskii, 2001; Oneto, Ghio, Ridella, & Anguita, 2015a), in fact:

$$\mathbb{E}_{\boldsymbol{\sigma}} \sup_{f\in\mathcal{F}} \frac{2}{n}\sum_{i=1}^{n} I(f(X_i), \sigma_i) = \mathbb{E}_{\boldsymbol{\sigma}} \sup_{f\in\mathcal{F}} \frac{1}{n}\sum_{i=1}^{n} \sigma_i f(X_i). \tag{8}$$

The Local Rademacher Complexity $\widehat{LR}_n(\mathcal{F}, r)$ (Bartlett et al., 2005; Koltchinskii, 2006), together with its expected value $LR_n(\mathcal{F}, r)$, is defined as:

$$\widehat{LR}_n(\mathcal{F}, r) = \widehat{R}_n\left(\left\{f : f \in \mathcal{F},\ \widehat{L}_n(f) \leq r\right\}\right), \tag{9}$$

$$LR_n(\mathcal{F}, r) = R_n\left(\{f : f \in \mathcal{F},\ L(f) \leq r\}\right). \tag{10}$$

The Local Rademacher Complexity improves over its global counterpart thanks to its ability of taking into account only those functions of the hypothesis class that will be most likely chosen by the learning procedure. This is due to the fact that in the definitions of Eqs. (9) and (10) the r parameter shrinks the hypothesis space by discarding the functions with large error. r is connected with the generalization ability of an $f \in \mathcal{F}$, as we will see in Theorem 3.5. Note, again, that the definitions of Local Rademacher Complexity of Eqs. (9) and (10) are in agreement with the ones that appeared in the recent literature (Bartlett et al., 2005; Koltchinskii, 2006; Oneto et al., 2015b), in fact:

$$\widehat{LR}_n(\mathcal{F}, r) = \mathbb{E}_{\boldsymbol{\sigma}} \sup_{f\in\left\{f : f\in\mathcal{F},\ \frac{1}{n}\sum_{i=1}^{n}[I(f(X_i),Y_i)]^2 \leq r\right\}} \times \frac{1}{n}\sum_{i=1}^{n} I(f(X_i), \sigma_i), \tag{11}$$

and

$$LR_n(\mathcal{F}, r) = \mathbb{E}_{X_1,\ldots,X_n}\mathbb{E}_{\boldsymbol{\sigma}} \sup_{f\in\left\{f : f\in\mathcal{F},\ \mathbb{E}_{(X,Y)}[I(f(X),Y)]^2 \leq r\right\}} \times \frac{1}{n}\sum_{i=1}^{n} I(f(X_i), \sigma_i), \tag{12}$$

since $I(f(X_i), Y_i) \in \{0, 1\}$ and, therefore, $[I(f(X_i), Y_i)]^2 = I(f(X_i), Y_i)$. Consequently, in this paper the definitions of Local Rademacher Complexity are not referred to $(\mathcal{F})$ but to $(I \circ \mathcal{F})$ since, as we will show in the next section, in this paper we are interested in relating the local complexity measures to the generalization error.

In the framework of the VC Theory, to the best of our knowledge, such approach has never been proposed: in fact, the VC Theory takes into account the whole hypothesis space.

In a recent preprint and unpublished article (Lei, Ding et al., 2015) an attempt of estimating the Local Rademacher Complexity via a Covering Number-based (Zhou, 2002) upper-bound has been made, but it still relies on the original work on Local Rademacher Complexity (Bartlett et al., 2005).

In this paper we propose, instead, a localized version of a complexity measure based on the VC Theory and show that it can be effectively exploited in learning theory for deriving a new generalization bound. This complexity measure extends the proposal of Vapnik (1998) by introducing the notion of localization in the traditional Vapnik's Statistical Learning Theory framework.

Let us localize the set of functions defined in Eq. (3) by introducing a constraint on the error, controlled by a parameter r:

$$\widehat{\mathcal{F}}_{(\mathcal{D}_n,r)} = \left\{\{f_1, \ldots, f_n\} : f \in \mathcal{F}, \widehat{L}_n(f) \leq r\right\}, \tag{13}$$

$$\mathcal{F}_{(\mathcal{D}_n,r)} = \left\{\{f_1, \ldots, f_n\} : f \in \mathcal{F}, L(f) \leq r\right\}, \tag{14}$$