Neural Networks 44 (2013) 107–111

Contents lists available at SciVerse ScienceDirect

Neural Networks

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

Neural networks letter

An improved analysis of the Rademacher data-dependent bound using its self bounding property

Luca Oneto *, Alessandro Ghio, Davide Anguita, Sandro Ridella

DITEN - University of Genova, Via Opera Pia 11A, I-16145 Genova, Italy

ARTICLE INFO

Article history:
Received 2 March 2012
Revised and accepted 22 March 2013

Keywords:
Error estimation
Data-dependent bounds
Rademacher complexity
Concentration of measure

ABSTRACT

The problem of assessing the performance of a classifier, in the finite-sample setting, has been addressed by Vapnik in his seminal work by using data-independent measures of complexity. Recently, several authors have addressed the same problem by proposing data-dependent measures, which tighten previous results by taking in account the actual data distribution. In this framework, we derive some data-dependent bounds on the generalization ability of a classifier by exploiting the Rademacher Complexity and recent concentration results: in addition of being appealing for practical purposes, as they exploit empirical quantities only, these bounds improve previously known results.

1. Introduction

The estimation of the performance of a model is a key issue in classification problems. Differently from asymptotic approaches, Vapnik's theory of Structural Risk Minimization (SRM) (Vapnik, 1998) allows one to target this problem in the finite-sample framework. In particular, SRM exploits the cardinality of the training set and the complexity of the hypothesis space for upper bounding the generalization error of a classifier, so to allow the identification of the optimal one.

The SRM proposes, for this purpose, a data-independent notion of complexity, which allows to derive several upper bounds of the generalization error of a model and are characterized by different convergence rates, with respect to the number of available samples. As an example, a first result can be obtained for a hypothesis space where none of the classifiers is characterized by a probability of error equal to zero (Vapnik, 1998, Section 4.1): in this *pessimistic* case, the estimated error decays as $O\left(n^{-1/2}\right)$. A second *optimistic* result, instead, considers the case where the hypothesis space consists of a finite number of models and at least one classifier has probability of error equal to zero (Vapnik, 1998, Section 4.2): in this case, the bound attains a convergence rate of $O\left(n^{-1}\right)$, thus a much quicker pace respect to the pessimistic case. Finally, a general bound can be derived, which lies between the optimistic and the pessimistic results (Vapnik, 1998, Section 4.3): the rate of convergence of this general bound is $O\left(n^{-\beta}\right)$, where the properties of the classifier affect the convergence rate, which can assume values between slow ($\beta = 1/2$) and fast convergence ($\beta = 1$).

More recently, alternative data-dependent bounds have been proposed, which can be studied by exploiting different approaches (Anguita, Ghio, Oneto, & Ridella, 2011a, 2011b; Anthony, 2008) and provide less conservative estimates of the generalization ability of a model. These approaches work by building an hypothesis space based on the actual available samples (Bartlett, Boucheron, & Lugosi, 2002; Bartlett, Bousquet, & Mendelson, 2005; Bartlett & Mendelson, 2003; Koltchinskii, 2001; Shawe-Taylor, Bartlett, Williamson, & Anthony, 1998), instead of considering a worst-case setting, like the classical SRM data-independent approach.

Several formulations of these bounds exist, which are characterized by different convergence rates. The bound proposed by Bartlett et al. (2005) and Shawe-Taylor et al. (1998), for example, decays at a fast pace and can be considered a data-dependent optimistic result; however, its finite-sample behavior for small n is very conservative (see for example Bartlett et al. (2005)) or the bounds need additional hypothesis (see for example Srebro, Sridharan, and Tewari (2010)) in order to be applied, representing a drawback, for example, when few tens of samples are available (Anguita, Ghio, Oneto, & Ridella, in press; Anguita, Ridella, Rivieccio, & Zunino, 2003; Braga-Neto & Dougherty, 2004; Magdon-Ismail, 2010; Pochet et al., 2005). On the contrary, the upper bound of the generalization error proposed by Bartlett and Mendelson (2003) and Koltchinskii (2001) is characterized by a slow convergence rate $O(n^{-1/2})$. For this reason, it can be considered a pessimistic result but, due to the fact that it is based on the well-known *Rademacher Complexity* statistical tool, it can be efficiently computed (Bartlett et al., 2002; Koltchinskii, 2001; Vapnik, 1999), making it very appealing.

In this paper, our objective is to exploit some relatively new concentration results by Boucheron, Lugosi, and Massart (2003,

* Corresponding author. Tel.: +39 0103532192.
E-mail addresses: Luca.Oneto@unige.it (L. Oneto), Alessandro.Ghio@unige.it (A. Ghio), Davide.Anguita@unige.it (D. Anguita), Sandro.Ridella@unige.it (S. Ridella).

0893-6080/$ – see front matter © 2013 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.neunet.2013.03.017

2000) and propose some improved and efficiently computable data-dependent bounds, which are sharper than the pessimistic data-dependent formulations proposed in Bartlett and Mendelson (2003) and Koltchinskii (2001). In particular, despite being characterized by the same slow asymptotical behavior $O(n^{-1/2})$, our bounds are sharper for small n, making them appealing in practical applications like, for example, error estimation of classifiers in the small sample setting (i.e. when small cardinality sets are available, e.g. see Anguita et al. (in press) and Braga-Neto and Dougherty (2004)). Furthermore, we show that these improved bounds contain explicitly both slow and fast convergence terms and the former ones are controlled by the classification performance of the model. In other words, we move a first step towards the identification of an effectively computable data-dependent general result, with varying convergence rate between $O(n^{-1/2})$ and $O(n^{-1})$.

For such purposes, we briefly recall the learning framework in Section 2 and, after having propaedeutically recalled in Section 3.1 the well-known results of Bartlett and Mendelson (2003) and Koltchinskii (2001), we improve them in Sections 3.2 and 3.3. Finally, in Section 4, we propose a comparison of the bounds in order to verify the sharpness of the proposed results.

2. The learning framework

We recall the standard probabilistic model in the framework of supervised learning, where the goal is to approximate a relationship between inputs, from a set $\mathcal{X}$, and outputs, from a set $\mathcal{Y}$. In this work, we target binary classification problems, as in the analysis of Kulkarni, Lugosi, and Venkatesh (1998), then we assume that $\mathcal{Y} \in \{-1, +1\}$. The relationship between inputs and outputs is encoded by a fixed, but unknown, probability distribution μ on $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$. The element $(\boldsymbol{x}, y) = \boldsymbol{z} \in \mathcal{Z}$ is defined as a labeled sample: the training phase consists in exploiting a set $(\boldsymbol{z}_1, \ldots, \boldsymbol{z}_n) \in \mathcal{Z}^n$ of labeled samples in a learning algorithm, which returns a function $h : \mathcal{X} \to \{-1, +1\}$ chosen in a fixed set $\mathcal{H}$ of possible hypotheses. The learning algorithm maps $(\boldsymbol{z}_1, \ldots, \boldsymbol{z}_n)$ to $\mathcal{H}$ and the accuracy in representing the hidden relationship μ is measured with reference to a binary loss function $\ell : \{-1, +1\} \times \mathcal{Y} \to \{0, 1\}$, which counts the number of misclassifications (e.g. $\ell(h(\boldsymbol{x}), y) = \frac{1-yh(\boldsymbol{x})}{2}$). For any $h \in \mathcal{H}$, we define the loss $L(h)$ as the expectation of $\ell(h(\boldsymbol{x}), y)$ with respect to μ, $L(h) = \mathbb{E}_\mu \ell(h(\boldsymbol{x}), y)$, where we assume that each labeled sample is independently generated according to μ.

3. Rademacher complexity bounds

In the following, we recall the results of Bartlett and Mendelson (2003) and Koltchinskii (2001) for the pessimistic data-dependent bound, as they allow us to introduce notation and concepts that will be useful for the successive analyses. The results for concentration inequalities can be retrieved in Boucheron et al. (2000) (for Self Bounding Functions) and McDiarmid (1989) (for Bounded Difference Functions).

3.1. The pessimistic data-dependent bound

As remarked in the previous section, $L(h)$ cannot be computed since μ is unknown. However, we can easily compute its empirical counterpart:

$$\hat{L}_n(h) = 1/n \sum_{i=1}^{n} \ell(h(\boldsymbol{x_i}), y_i). \tag{1}$$

As we do not know in advance which model h will be chosen by the learning algorithm, a classical approach consists in studying the uniform deviation of $\hat{L}_n(h)$ from $L(h)$, respect to all possible models (Anthony, 2008; Bartlett et al., 2002; Bartlett & Mendelson, 2003; Koltchinskii, 2001; Vapnik, 1999).

Definition 1. The uniform deviation is defined as

$$\hat{\delta}_n(\mathcal{H}) = \sup_{h \in \mathcal{H}} \left[L(h) - \hat{L}_n(h) \right], \tag{2}$$

and, therefore, the following inequality holds:

$$L(h) - \hat{L}_n(h) \leq \hat{\delta}_n(\mathcal{H}), \tag{3}$$

which allows to upper bound $L(h)$ in terms of its empirical counterparts.

For studying this quantity, Bartlett and Mendelson (2003) suggested to use the *Rademacher Complexity*, a well-known statistical tool for measuring the complexity of a class of functions.

Definition 2. The Rademacher Complexity can be defined as

$$\hat{\mathcal{R}}_n(\mathcal{H}) = \mathbb{E}_{\boldsymbol{\sigma}} \sup_{h \in \mathcal{H}} \frac{2}{n} \sum_{i=1}^{n} \sigma_i \ell(h(\boldsymbol{x}_i), y_i), \tag{4}$$

where $\sigma_1, \ldots, \sigma_n$ are independent uniform $\{\pm 1\}$-valued random variables.

In our case, as we are targeting binary classification problems, the following lemma can be trivially proven.

Lemma 3.1. *In the case of binary classification, the Rademacher Complexity can be written as*

$$\begin{aligned} \hat{\mathcal{R}}_n(\mathcal{H}) &= \mathbb{E}_{\boldsymbol{\sigma}} \sup_{h \in \mathcal{H}} \frac{2}{n} \sum_{i=1}^{n} \sigma_i \ell(h(\boldsymbol{x}_i), y_i) \\ &= \mathbb{E}_{\boldsymbol{\sigma}} \sup_{h \in \mathcal{H}} \frac{1}{n} \sum_{i=1}^{n} \sigma_i h(\boldsymbol{x}_i). \end{aligned} \tag{5}$$

The uniform deviation is a random variable, so it is possible to study its expected value $\mathbb{E}_\mu \hat{\delta}_n(\mathcal{H})$ and relate it to the Rademacher Complexity of the class (Bartlett et al., 2002; Bartlett & Mendelson, 2003; Giné & Zinn, 1984).

Theorem 3.2. *The Rademacher Complexity and the uniform deviation satisfy the following property:*

$$\mathbb{E}_\mu \hat{\delta}_n(\mathcal{H}) \leq \mathbb{E}_\mu \hat{\mathcal{R}}_n(\mathcal{H}). \tag{6}$$

Thanks to the concentration results described in McDiarmid (1989), it is easy to prove that, with high probability, both the Rademacher Complexity and the uniform deviation are sharply concentrated around their means, as shown by Bartlett and Mendelson (2003). Thus the following, now classic, result can be derived (Bartlett & Mendelson, 2003):

Theorem 3.3. *With probability* $(1 - \delta)$*:*

$$L(h) - \hat{L}_n(h) \leq \mathbb{E}_\mu \hat{\mathcal{R}}_n(\mathcal{H}) + \sqrt{\frac{\log\left(\frac{1}{\delta}\right)}{2n}}, \quad \forall h \in \mathcal{H}. \tag{7}$$

Note that this bound cannot be computed in practice, because it requires that μ is known. However, it is possible to derive its fully empirical version (Bartlett & Mendelson, 2003; Koltchinskii, 2001), which clearly shows the $O(n^{-1/2})$ rate of convergence.

Theorem 3.4. *With probability* $(1 - \delta)$ *we have that:*

$$L(h) - \hat{L}_n(h) \leq \hat{\mathcal{R}}_n(\mathcal{H}) + 3\sqrt{\frac{\log\left(\frac{2}{\delta}\right)}{2n}}, \quad \forall h \in \mathcal{H}. \tag{8}$$