Pattern Recognition Letters 37 (2014) 210–219

Contents lists available at SciVerse ScienceDirect

Pattern Recognition Letters

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

Unlabeled patterns to tighten Rademacher complexity error bounds for kernel classifiers

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

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

ARTICLE INFO

Article history:
Available online 10 May 2013

Keywords:
Support vector machine
Rademacher complexity
Structural risk minimization
Error estimation
Model selection

ABSTRACT

We derive in this work new upper bounds for estimating the generalization error of kernel classifiers, that is the misclassification rate that the models will perform on new and previously unseen data. Though this paper is more targeted towards the error estimation topic, the generalization error can be obviously exploited, in practice, for model selection purposes as well. The derived bounds are based on Rademacher complexity and result to be particularly useful when a set of unlabeled samples are available, in addition to the (labeled) training examples: we will show that, by exploiting further unlabeled patterns, the confidence term of the conventional Rademacher complexity bound can be reduced by a factor of three. Moreover, the availability of unlabeled examples allows also to obtain further improvements by building localized versions of the hypothesis class containing the optimal classifier.

1. Introduction

Getting a deep insight into the factors that influence the performance of a statistical procedure is a basic step to improve models reliability and effectiveness. When focussing on machine learning approaches to pattern classification, one of the most explored procedures aims at solving the well-known *error estimation* problem, targeting the estimation of the generalization error of a classifier, i.e. the misclassification rate that the predictor will perform on new future samples. Error estimation procedures are straightforwardly linked to the *model selection* problem, whose goal is the choice (based on the estimated generalization error) of the optimal classifier from a set of possible models, namely the *hypothesis space*.

Several approaches have been proposed for such purposes (e.g. Vapnik and Chervonenkis, 1971; Bartlett and Mendelson, 2003; Bartlett et al., 2005; Bousquet and Elisseeff, 2002), which allow to provide upper-bounds of the generalization performance of a classifier: these quantities can be used as an index for comparing different models and choosing the best performing predictor during the model selection phase. These upper-bounds usually consist of three main terms, i.e.:

1. The empirical error of the classifier, performed by the model on the training data, used for the predictor learning;
2. A bias, where the complexity of the hypothesis space, where the classifier is picked-up from, is taken into account;
3. Finally, a confidence term, usually independent of the hypothesis space and the chosen classifier, which only depends on the user-defined confidence value of the bound and on the cardinality of the training set.

The objective of these approaches is to investigate the finite sample behavior of a model instead of the asymptotic one: despite being appealing for real-world problems, their practical applicability has been questioned for a long time. Concerning finite sample bounds, one of the most recent and effective approaches relies on the Rademacher complexity (RC), a powerful statistical tool that has been investigated throughout the last years (Bartlett et al., 2002; Anguita et al., 2012; Anguita et al., 2011c), for which a practical procedure targeting the use of RC bounds to model selection and error estimation of kernel classifiers has been recently proposed in Anguita et al. (2012).

The RC approach, in particular, showed to be more effective than traditional techniques (e.g. Arlot and Celisse, 2010; Efron and Tibshirani, 1993) when applied to the small-sample regime (Anguita et al., 2011b; Bartlett et al., 2002; Anguita et al., 2011c), i.e. problems where the cardinality of the labeled training set is comparable to or (even remarkably) lower than the dimensionality of the samples. In this work we show how RC bounds can be further improved by exploiting eventually available extra-knowledge on the phenomenon to be modeled: this additional information has the form of unlabeled data, that are often available in real-world pattern classification problems, as also confirmed by the growing interest in the semi-supervised learning framework (Bennett and Demiriz, 1999; Chapelle et al., 2010). On the opposite of the approaches, proposed in this latter framework, we do not focus

* Corresponding author. Tel.: +39 (0)10 3532192; fax: +39 (0)10 3532897.
E-mail addresses: davide.anguita@unige.it (D. Anguita), alessandro.ghio@unige.it (A. Ghio), luca.oneto@unige.it (L. Oneto), sandro.ridella@unige.it (S. Ridella).

0167-8655/$ - see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.patrec.2013.04.027

on how exploiting unlabeled data for training a model, as we move the spotlights on the error estimation step, by properly modifying the RC theory so that it can exploit the unlabeled extra-knowledge. In particular, we first show how a new RC bound can be derived, which contemplates both labeled and unlabeled data, allowing to reduce the confidence term; moreover, we also propose a method, based on the previous work of Anguita et al. (2012), allowing to use unlabeled data for selecting a more effective, problem-dependent and local hypothesis space, resulting in a much sharper and accurate bound.

2. Theoretical framework and results

We consider the following pattern classification problem: based on a random observation of $X \in \mathcal{X} \subseteq \mathbb{R}^d$ one has to estimate $Y \in \mathcal{Y} \subseteq \{-1,+1\}$ by choosing a suitable prediction rule $f : X \to [-1,+1]$. The generalization error $L(f) = \mathbb{E}_{\{\mathcal{X},\mathcal{Y}\}}\ell(f(X),Y)$ of f is defined through a bounded loss function $\ell(f(X),Y) : [-1,+1]\times\mathcal{Y} \to [0,1]$. Let $\mathcal{D}_{n_l} : \left\{(X_1^l,Y_1^l),\ldots,(X_{n_l}^l,Y_{n_l}^l)\right\}$ be a set of independent and identically distributed (i.i.d.) labeled samples and $\mathcal{D}_{n_u} : \{(X_1^u),\ldots,(X_{n_u}^u)\}$ a set of i.i.d. unlabeled patterns, originated by the same distribution $P(\mathcal{X},\mathcal{Y})$.

As $P(\mathcal{X},\mathcal{Y})$ is obviously unknown, $L(f)$ cannot be directly computed. We can compute, instead, the empirical estimation of $L(f)$ on the set of labeled data:

$$L_{n_l}(f) = \frac{1}{n_l}\sum_{i=1}^{n_l}\ell(f(X_i^l),Y_i^l) \tag{1}$$

which, however, cannot be safely used for error estimation purposes as $L_{n_l}(f)$ is a clearly optimistically-biased estimation of the generalization error. Our objective is to derive a statistical sound upper bound of $L(f)$, by taking into account the information embedded in both $\mathcal{D}_{n_l}$ and $\mathcal{D}_{n_u}$.

In the well-known framework of *Structural Risk Minimization* (SRM) (Vapnik, 2000), we have to define an infinite sequence of hypothesis spaces of increasing complexity $\{\mathcal{F}_i,\ i = 1,2,\ldots\}$: thus, we have to choose the most suitable function space $\mathcal{F}_i$ and, obviously, the model $f^* \in \mathcal{F}_i$ that is characterized by the best generalization ability within $\mathcal{F}_i$. As the true data distribution, originating the data, is unknown, it is only possible to state that:

$$\left\{L(f) - L_{n_l}(f)\right\}_{f\in\mathcal{F}_i} \leqslant \sup_{f\in\mathcal{F}_i}\left\{L(f) - L_{n_l}(f)\right\} \tag{2}$$

or, equivalently:

$$L(f) \leqslant L_{n_l}(f) + \sup_{f\in\mathcal{F}_i}\left\{L(f) - L_{n_l}(f)\right\}, \quad \forall f \in \mathcal{F}_i. \tag{3}$$

In this framework, according to the SRM procedure, the following function space and the corresponding optimal classifier are chosen:

$$f^*,\mathcal{F}^* : \arg\min_{\mathcal{F}_i\in\{\mathcal{F}_1,\mathcal{F}_2,\ldots\}}\left[\min_{f\in\mathcal{F}_i} L_{n_l}(f)_{f\in\mathcal{F}_i} + \sup_{f\in\mathcal{F}_i}\left\{L(f) - L_{n_l}(f)\right\}\right]. \tag{4}$$

The *generalization bias* ($\sup_{f\in\mathcal{F}_i}\{L(f) - L_{n_l}(f)\}$) is a random variable, thus it is possible to statistically analyze it and derive a bound which holds with a user-defined probability, e.g. as shown in Bartlett et al. (2002).

In our analysis, in particular, we restrict to two types of suitable prediction rules, since their associated loss functions, namely the *hard loss* $\ell_H(f_H(\boldsymbol{x}),y)$ and the *soft (or ramp) loss* (Collobert et al., 2006) $\ell_S(f_S(\boldsymbol{x}),y)$, are bounded ($[0,1]$) and symmetric ($\ell(f(\boldsymbol{x}),y) = 1 - \ell(f(\boldsymbol{x}),-y)$):

$$f_H(\boldsymbol{x}) = \operatorname{sign}(\boldsymbol{w}^T\boldsymbol{\phi}(\boldsymbol{x}) + b), \tag{5}$$

$$\ell_H(f_H(\boldsymbol{x}),y) = \frac{1 - yf_H(\boldsymbol{x})}{2}, \tag{6}$$

$$f_S(\boldsymbol{x}) = \begin{cases}\min(1,\boldsymbol{w}^T\boldsymbol{\phi}(\boldsymbol{x}) + b) & \text{if} \quad \boldsymbol{w}^T\boldsymbol{\phi}(\boldsymbol{x}) + b > 0,\\ \max(-1,\boldsymbol{w}^T\boldsymbol{\phi}(\boldsymbol{x}) + b) & \text{if} \quad \boldsymbol{w}^T\boldsymbol{\phi}(\boldsymbol{x}) + b \leqslant 0,\end{cases} \tag{7}$$

$$\ell_S(f_S(\boldsymbol{x}),y) = \frac{1 - yf_S(\boldsymbol{x})}{2}, \tag{8}$$

where $\boldsymbol{\phi}(\cdot) : \mathbb{R}^d \to \mathbb{R}^D$ with $D \gg d$, $\boldsymbol{w} \in \mathbb{R}^D$ and $b \in \mathbb{R}$. The function $\boldsymbol{\phi}(\cdot)$ is introduced to allow for a later exploitation of kernels; however, we will focus only on the linear case in this paper, as it simplifies the discussion: the usual non-linear formulation can be easily derived by applying the well-known kernel trick (Shawe-Taylor and Cristianini, 2000).

We recall the definition of *Rademacher complexity* (*RC*) of a class of functions $\mathcal{F}$:

$$\hat{\mathcal{R}}_{n_l}(\mathcal{F}) = \mathbb{E}_\sigma\sup_{f\in\mathcal{F}}\frac{2}{n_l}\sum_{i=1}^{n_l}\sigma_i\ell(f(\boldsymbol{x}_i),y_i) = \mathbb{E}_\sigma\sup_{f\in\mathcal{F}}\frac{1}{n_l}\sum_{i=1}^{n_l}\sigma_i f(\boldsymbol{x}_i) \tag{9}$$

where $\sigma_1,\ldots,\sigma_{n_l}$ are n_l independent Rademacher random variables, i.e. independent random variables for which $\mathbb{P}(\sigma_i = +1) = \mathbb{P}(\sigma_i = -1) = 1/2$. It is worth underlining that the last equality holds if one of the losses introduced above is exploited, as it is valid only for bounded and symmetric functions. The quantity in Eq. (9) is a computable realization of the expected Rademacher complexity $\mathcal{R}(\mathcal{F}) = \mathbb{E}_{(\mathcal{X},\mathcal{Y})}\hat{\mathcal{R}}_{n_l}(\mathcal{F})$. The most renewed result in Rademacher complexity theory states that (Bartlett and Mendelson, 2003):

$$L(f)_{f\in\mathcal{F}} \leqslant L_{n_l}(f)_{f\in\mathcal{F}} + \hat{\mathcal{R}}_{n_l}(\mathcal{F}) + 3\sqrt{\frac{\log\frac{2}{\delta}}{2n_l}} \tag{10}$$

which holds with probability $(1-\delta)$ and can be exploited in the SRM framework (see Eq. (4)). Note that the previous bound does not allow to exploit the information included in the unlabeled samples: our first result will thus allow to contemplate $\mathcal{D}_{n_u}$ in the estimation of the generalization performance of f.

2.1. Exploiting unlabeled samples for reducing the confidence term

We can safely assume that the number of unlabeled samples is larger than the cardinality of the labelled training patterns. Thus, we can split the unlabeled data in blocks of similar size by defining the quantity $m = \lfloor n_u/n_l \rfloor + 1$, so to create a ghost sample D'_{mn_l} consisting of mn_l patterns. Then, we can upper bound the expected generalization bias as follows[1]:

$$\begin{aligned}
\mathbb{E}_{\{\mathcal{X},\mathcal{Y}\}}\sup_{f\in\mathcal{F}}\left\{L(f) - L_{n_l}(f)\right\} &= \mathbb{E}_{\{\mathcal{X},\mathcal{Y}\}}\sup_{f\in\mathcal{F}}\left[\mathbb{E}_{\{\mathcal{X}',\mathcal{Y}'\}}\left[\frac{1}{m}\sum_{i=1}^{m}\frac{1}{n_l}\sum_{k=(i-1)\cdot n_l+1}^{i\cdot n_l}\ell'_k\right] - \frac{1}{n_l}\sum_{i=1}^{n_l}\ell_i\right]\\
&\leqslant \mathbb{E}_{\{\mathcal{X},\mathcal{Y}\}}\mathbb{E}_{\{\mathcal{X}',\mathcal{Y}'\}}\\
&\quad\times\frac{1}{m}\sum_{i=1}^{m}\sup_{f\in\mathcal{F}}\left[\frac{1}{n_l}\sum_{k=(i-1)\cdot n_l+1}^{i\cdot n_l}\left(\ell'_k - \ell_{|k|_{n_l}}\right)\right]\\
&= \mathbb{E}_{\{\mathcal{X},\mathcal{Y}\}}\mathbb{E}_{\{\mathcal{X}',\mathcal{Y}'\}}\mathbb{E}_\sigma\\
&\quad\times\frac{1}{m}\sum_{i=1}^{m}\sup_{f\in\mathcal{F}}\left[\frac{1}{n_l}\sum_{k=(i-1)\cdot n_l+1}^{i\cdot n_l}\sigma_{|k|_{n_l}}\left[\ell'_k - \ell_{|k|_{n_l}}\right]\right]\\
&\leqslant \mathbb{E}_{\{\mathcal{X},\mathcal{Y}\}}\mathbb{E}_\sigma\frac{1}{m}\sum_{i=1}^{m}\sup_{f\in\mathcal{F}}\left[\frac{2}{n_l}\sum_{k=(i-1)\cdot n_l+1}^{i\cdot n_l}\sigma_{|k|_{n_l}}\ell_k\right]\\
&= \mathbb{E}_{\{\mathcal{X},\mathcal{Y}\}}\frac{1}{m}\sum_{i=1}^{m}\hat{\mathcal{R}}^i_{n_l}(\mathcal{F})
\end{aligned} \tag{11}$$

where $|k|_{n_l} = (k-1) \bmod (n_l) + 1$. The last quantity, which we define the *expected extended Rademacher complexity*

[1] We define $\ell(f(\boldsymbol{x}_i),y_i) \equiv \ell_i$ for notational simplicity.