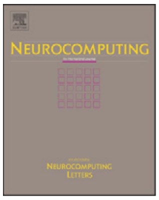
Neurocomputing 74 (2011) 1436–1443

Contents lists available at ScienceDirect

Neurocomputing

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



Maximal Discrepancy for Support Vector Machines

Davide Anguita, Alessandro Ghio *, Sandro Ridella

Department of Biophysical and Electronic Engineering, University of Genova, Via Opera Pia 11a, I-16145 Genova, Italy

ARTICLE INFO

Available online 19 February 2011

Keywords:
Maximal discrepancy (MD)
Error estimation
Support vector machine (SVM)

ABSTRACT

The Maximal Discrepancy (MD) is a powerful statistical method, which has been proposed for model selection and error estimation in classification problems. This approach is particularly attractive when dealing with small sample problems, since it avoids the use of a separate validation set. Unfortunately, the MD method requires a bounded loss function, which is usually avoided by most learning algorithms, including the Support Vector Machine (SVM), because it gives rise to a non-convex optimization problem. We derive in this work a new approach for rigorously applying the MD technique to the error estimation of the SVM and, at the same time, preserving the original SVM framework.

1. Introduction

We propose in this work a new method to upper bound the generalization error of Support Vector Machines (SVMs) [1–3] using a data-dependent bound, based on Maximal Discrepancy (MD) [4]. Several data-dependent methods have been recently proposed in the literature: they use different approaches based on worst-case or empirical fat-shattering [5,6], bootstrap penalties [7] and Rademacher or MD complexities [4,8–10], but their practical application to the usual formulation of SVM is not trivial. In particular, when adopting Rademacher or MD-based approaches, one of the main problems lies in the fact that these methods require the use of a Lipschitz and bounded loss function, which is not the case for the conventional SVM. The last requirement, in fact, gives rise to a non-convex optimization problem, for which some solutions have been proposed in the literature, but they depart from the traditional SVM framework and require ad-hoc learning procedures [11–16].

We propose to solve this issue through a new learning procedure, which allows us to deal with a convex optimization problem and, furthermore, to rigorously apply the MD method to the SVM classifier. The price to pay for our approach is a slight increase of the generalization error estimate, but the method can be easily applied in practice.

Our scenario is a typical learning problem where a set of i.i.d. patterns $D_l = \{(\boldsymbol{x}_1,y_1),\ldots,(\boldsymbol{x}_l,y_l)\}$ is available, sampled from the unknown distributions $P(\boldsymbol{x})$ and $P(y|\boldsymbol{x})$, with $\boldsymbol{x}_i \in \mathbb{R}^n$ and $y_i \in \mathcal{Y} = \{-1,+1\}$.

A *prediction rule* is a function $f : \mathbb{R}^n \rightarrow \mathcal{Y}_f \subseteq \mathbb{R}$, chosen from a set of functions $\mathcal{F}$, whose empirical error (i.e. the error rate on the dataset D_l) can be written as

$$\nu(f) = \frac{1}{l}\sum_{i=1}^{l} L(f(\boldsymbol{x}_i),y_i), \tag{1}$$

where $L : \mathcal{Y}_f \times \mathcal{Y} \rightarrow [0,1]$ is a *loss function* that weights the mispredicted samples.[1]

A typical example is the following loss function, which assigns a unit weight to misclassified samples and a zero value to correctly classified ones:

$$L_H(f(\boldsymbol{x}_i),y_i) = \begin{cases} 0 & \text{if } y_i f(\boldsymbol{x}_i) > 0, \\ 1 & \text{if } y_i f(\boldsymbol{x}_i) \le 0. \end{cases} \tag{2}$$

Unfortunately, the use of a *hard* loss function makes the optimization problem intractable, therefore, a *soft* version is used instead, which is Lipschitz continuous and can assume any value in the range [0,1].

We propose to use the following soft loss function:

$$L_S(f(\boldsymbol{x}_i),y_i) = \begin{cases} 0 & \text{if } y_i f(\boldsymbol{x}_i) \ge 1, \\ (1-y_i f(\boldsymbol{x}_i))/2 & \text{if } -1 \le y_i f(\boldsymbol{x}_i) \le 1, \\ 1 & \text{if } y_i f(\boldsymbol{x}_i) \le -1. \end{cases} \tag{3}$$

Differently from other proposals [11,12,15], this loss function assigns a weight of $\frac{1}{2}$ to the samples that are exactly on the separating surface and the extreme values 0 or 1 are assigned to the patterns that lie outside the margin $|y_i f(\boldsymbol{x}_i)| \ge 1$. The value given by Eq. (3) cannot be interpreted directly as a conditional probability, which, in any case, cannot be unambiguously estimated in sparse classifiers [18]; nevertheless, it is a real-valued

* Corresponding author. Tel.: +39 010 3532192; fax: +39 010 3532897.
E-mail addresses: Davide.Anguita@unige.it (D. Anguita), Alessandro.Ghio@unige.it (A. Ghio), Sandro.Ridella@unige.it (S. Ridella).

[1] In this paper, we adopt the notation used in [4,17].

0925-2312/$ - see front matter © 2011 Elsevier B.V. All rights reserved.
doi:10.1016/j.neucom.2010.12.009

indicator of the amount of misclassification and, furthermore, can be used to upper bound the number of misclassified samples, because $L_H \le 2L_S$.

In order to predict the generalization ability of the classifier, we are interested in the estimation of the *generalization error* $\pi(f)$, which is defined as

$$\pi(f) = \mathbb{E}_{(\boldsymbol{x},y)} L(f(\boldsymbol{x}),y), \tag{4}$$

where $\mathbb{E}_{(\boldsymbol{x},y)}$ is the expectation with respect to the probability distribution of the data $\boldsymbol{x}$ and the labels y. It is well known that $\nu(f)$ usually underestimates $\pi(f)$, but the last one cannot be computed because the probability distribution generating the data is unknown. Several well-known results in learning theory (see, for example, [19]) show that a key quantity for estimating the generalization ability of a classifier is the supremum of the *generalization bias* with respect to the considered class of functions $\mathcal{F}$, $\sup_{f\in\mathcal{F}}[\pi(f)-\nu(f)]$. This is a random variable depending on the data and $\mathcal{F}$, for which the following relation is always true

$$\pi(f) \le \nu(f) + \sup_{g\in\mathcal{F}}[\pi(g)-\nu(g)], \tag{5}$$

and which can be analyzed through the use of maximal discrepancy.

In the following section we will describe the MD method and in Section 3 we will summarize the SVM algorithm and the critical issues that must be addressed for applying the MD method to this classifier. In Section 4, we will detail our procedure, which allows to apply the MD method to the SVM in a rigorous way and, in Section 5, we will show some experimental results on a well-known benchmarking classification problem. Finally, in Section 6, we address some open problems related to the looseness of the bound, which can be a starting point for further research on this issue.

2. The Maximal Discrepancy of a classifier

Let us randomly split D_l in two halves and define the two following empirical errors:

$$\nu^{(1)}(f) = \frac{2}{l}\sum_{i=1}^{l/2} L(f(\boldsymbol{x}_i),y_i), \tag{6}$$

$$\nu^{(2)}(f) = \frac{2}{l}\sum_{i=l/2+1}^{l} L(f(\boldsymbol{x}_i),y_i). \tag{7}$$

The Maximal Discrepancy is defined as

$$MD = \max_{f\in\mathcal{F}}(\nu^{(1)}(f)-\nu^{(2)}(f)), \tag{8}$$

where $\mathcal{F}$ is the class of functions from which the classifier will be selected.

The following theorem can be used to derive an upper-bound of the generalization error in terms of MD [4,8]:

Theorem 1. *Given a dataset $D_l = \{(\boldsymbol{x}_1,y_1),\ldots,(\boldsymbol{x}_l,y_l)\}$, where $\boldsymbol{x}_i \in \mathbb{R}^n$ and $y_i \in \mathcal{Y}$, i.i.d. with unknown distributions $P(\boldsymbol{x})$ and $P(y|\boldsymbol{x})$, given a class of functions $\mathcal{F}$ and a bounded loss function $L : \mathcal{Y}_f \times \mathcal{Y} \to [0,1]$, then*

$$\mathbb{P}\left[\sup_{f\in\mathcal{F}}[\pi(f)-\nu(f)] \ge MD + \varepsilon\right] \le 2\exp\left(\frac{-2l\varepsilon^2}{9}\right). \tag{9}$$

Proof. The proof is similar to those of Theorem 9 in [4] and Theorem 8 in [8], in which the authors exploit the hard loss function, and is mainly based on an application of McDiarmid's inequality [20]. By considering a *ghost* sample $D'_l = \{\boldsymbol{x}'_i, y'_i\}$, composed of l patterns, the following upper bound holds:

$$\mathbb{E}_{(\boldsymbol{x},y)}\sup_{f\in\mathcal{F}}[\pi(f)-\nu(f)] = \mathbb{E}_{(\boldsymbol{x},y)}\sup_{f\in\mathcal{F}}\{\mathbb{E}_{(\boldsymbol{x}',y')}[\nu'(f)]-\nu(f)\} \tag{10}$$

$$= \mathbb{E}_{(\boldsymbol{x},y)}\sup_{f\in\mathcal{F}}\{\mathbb{E}_{(\boldsymbol{x}',y')}[\nu'(f)-\nu(f)]\}$$

$$\le \mathbb{E}_{(\boldsymbol{x},y)}\mathbb{E}_{(\boldsymbol{x}',y')}\left\{\max_{f\in\mathcal{F}}[\nu'(f)-\nu(f)]\right\}, \tag{11}$$

where $\nu'(f)$ is the empirical error over the ghost set and $\pi(f) = \mathbb{E}_{(\boldsymbol{x}',y')}[\nu'(f)]$. We exploit the definition of empirical error and, by defining $L_i = L(f(\boldsymbol{x}_i),y_i)$ and $L'_i = L(f(\boldsymbol{x}'_i),y'_i)$, we have

$$\begin{aligned}
&\mathbb{E}_{(\boldsymbol{x},y)}\mathbb{E}_{(\boldsymbol{x}',y')}\left\{\max_{f\in\mathcal{F}}[\nu'(f)-\nu(f)]\right\}\\
&\quad= \mathbb{E}_{(\boldsymbol{x},y)}\mathbb{E}_{(\boldsymbol{x}',y')}\left\{\max_{f\in\mathcal{F}}\left[\frac{1}{l}\sum_{i=1}^{l}L'_i - \frac{1}{l}\sum_{i=1}^{l}L_i\right]\right\}\\
&\quad\le \frac{1}{l}\mathbb{E}_{(\boldsymbol{x},y)}\mathbb{E}_{(\boldsymbol{x}',y')}\left\{\max_{f\in\mathcal{F}}\left[\sum_{i=1}^{l/2}(L'_i-L_i)\right]\right\}+\\
&\quad+\frac{1}{l}\mathbb{E}_{(\boldsymbol{x},y)}\mathbb{E}_{(\boldsymbol{x}',y')}\left\{\max_{f\in\mathcal{F}}\left[\sum_{i=l/2+1}^{l}(L'_i-L_i)\right]\right\}\\
&\quad= \frac{2}{l}\mathbb{E}_{(\boldsymbol{x},y)}\left\{\max_{f\in\mathcal{F}}\left[\sum_{i=1}^{l/2}(L'_i-L_i)\right]\right\} = \mathbb{E}_{(\boldsymbol{x},y)}\{MD\}.
\end{aligned} \tag{12}$$

Then, by considering Eqs. (10) and (12):

$$\mathbb{E}_{(\boldsymbol{x},y)}\sup_{f\in\mathcal{F}}[\pi(f)-\nu(f)] \le \mathbb{E}_{(\boldsymbol{x},y)}\{MD\}. \tag{13}$$

By applying McDiarmid's inequality [20] to the first term of Eq. (13):

$$P\left[\sup_{f\in\mathcal{F}}[\pi(f)-\nu(f)] \ge \mathbb{E}_{(\boldsymbol{x},y)}\sup_{f\in\mathcal{F}}[\pi(f)-\nu(f)]+\varepsilon\right] \le \exp[-2l\varepsilon^2], \tag{14}$$

since the loss $L(\cdot,\cdot)$ is bounded. By combining Eqs. (13) and (14), we obtain

$$\begin{aligned}
&P\left[\sup_{f\in\mathcal{F}}[\pi(f)-\nu(f)] \ge \mathbb{E}_{(\boldsymbol{x},y)}MD+\varepsilon\right]\\
&\quad\le P\left[\sup_{f\in\mathcal{F}}[\pi(f)-\nu(f)] \ge \mathbb{E}_{(\boldsymbol{x},y)}\sup_{f\in\mathcal{F}}[\pi(f)-\nu(f)]+\varepsilon\right] \le \exp[-2l\varepsilon^2].
\end{aligned} \tag{15}$$

We are interested in evaluating the generalization error using *MD*; then, we have

$$\begin{aligned}
&P\left[\sup_{f\in\mathcal{F}}[\pi(f)-\nu(f)] \ge MD+\varepsilon\right]\\
&\quad\le P\left[\sup_{f\in\mathcal{F}}[\pi(f)-\nu(f)] \ge \mathbb{E}_{(\boldsymbol{x},y)}\sup_{f\in\mathcal{F}}[\pi(f)-\nu(f)]+\frac{\varepsilon}{3}\right]\\
&\quad+P\left[\mathbb{E}_{(\boldsymbol{x},y)}\sup_{f\in\mathcal{F}}[\pi(f)-\nu(f)] \ge MD+\frac{2\varepsilon}{3}\right].
\end{aligned} \tag{16}$$

By considering Eq. (13), we have

$$\begin{aligned}
&P\left[\sup_{f\in\mathcal{F}}[\pi(f)-\nu(f)] \ge MD+\varepsilon\right]\\
&\quad\le P\left[\sup_{f\in\mathcal{F}}[\pi(f)-\nu(f)] \ge \mathbb{E}_{(\boldsymbol{x},y)}\sup_{f\in\mathcal{F}}[\pi(f)-\nu(f)]+\frac{\varepsilon}{3}\right]\\
&\quad+P\left[\mathbb{E}_{(\boldsymbol{x},y)}\{MD\} \ge MD+\frac{2\varepsilon}{3}\right] \le \exp\left[\frac{-2l\varepsilon^2}{9}\right]+\exp\left[\frac{-2l\varepsilon^2}{9}\right]\\
&\quad= 2\exp\left[\frac{-2l\varepsilon^2}{9}\right],
\end{aligned} \tag{17}$$