ESANN 2011 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 27-29 April 2011, i6doc.com publ., ISBN 978-2-87419-044-5.
Available from http://www.i6doc.com/en/livre/?GCOI=28001100817300.

Maximal Discrepancy Vs. Rademacher Complexity for Error Estimation

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

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

Abstract. The Maximal Discrepancy and the Rademacher Complexity are powerful statistical tools that can be exploited to obtain reliable, albeit not tight, upper bounds of the generalization error of a classifier. We study the different behavior of the two methods when applied to linear classifiers and suggest a practical procedure to tighten the bounds. The resulting generalization estimation can be succesfully used for classifier model selection.

1 Introduction

When targeting small–sample classification problems, where the cardinality of the training set is very small, typical *hold–out* techniques, like Cross Validation [1], can be unreliable [2]. These methods, in fact, waste some data for estimating the classification error by building a separate test set, so further reducing the size of the training set and the reliability of the classifier itself. *In–sample* techniques, instead, use the entire learning set both for training the classifier and estimating its generalization error [3, 4, 5, 6], so that their use in the small sample setting is very appealing. In addition, this estimation can also be used for model selection purposes, when the learning procedure requires the tuning of additional hyper–parameters. Hold–out techniques, instead, require to resort to nested procedures, which remove even more data from the training set to build both a validation set, for model selection purposes and a test set, for error estimation of the selected classifier.

Unfortunately, in–sample techniques are seldomly used in practice because their application to state–of–the–art classification algorithms, like the Support Vector Machine [3], is not trivial. Recently, however, some effective approaches have been proposed [7, 8, 9], which make them competitive with hold–out methods. The two best–known in–sample techniques are the *Maximal Discrepancy* (MD) [5] and the *Rademacher Complexity* (RC) [6]. Our purpose is to verify if and under which conditions MD outperforms RC, or vice versa, in estimating the true error of the classifier. As the estimation of the error provided by in–sample techniques is sometimes too loose to be of any practical use, we propose an heuristic procedure for tightening the bounds, which exploits some recent results [10]. A positive outcome of this procedure is to improve the applicability of the MD and RC methods to the model selection of classifiers.

ESANN 2011 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 27-29 April 2011, i6doc.com publ., ISBN 978-2-87419-044-5. Available from http://www.i6doc.com/en/livre/?GCOI=28001100817300.

2 Classification and error estimation: Maximal Discrepancy and Rademacher Complexity

Let us consider a set X of n i.i.d. patterns $(\boldsymbol{x}_i, y_i)$, with $\boldsymbol{x}_i \in \ ^d$ and $y_i \in \mathcal{Y} = \{\pm 1\}$, sampled from a distribution $\mathcal{P}(\boldsymbol{x}, y)$. Given a linear classifier $f(\boldsymbol{x}) = \boldsymbol{w} \cdot \boldsymbol{x} + b$, $f : \ ^d \rightarrow \mathcal{Y}_f \subseteq \ $, we can easily compute its empirical error $\hat{L}_n(f) = \frac{1}{n}\sum_{i=1}^{n} \ell(f(\boldsymbol{x}_i), y_i)$ on the set X, where $\ell(\cdot,\cdot)$ is a *loss function*. Our objective is to find a good and reliable estimation of the generalization error $L(f) = \ _{(\boldsymbol{x},y)}\, \ell\left(f(\boldsymbol{x}_i), y_i\right)$, which cannot be directly computed as $\mathcal{P}(\boldsymbol{x}, y)$ is unknown. The empirical error is of little help in this respect because it is well-known that $\hat{L}_n(f)$ usually underestimates $L(f)$. In particular, the function $f^* = \arg\min_{f\in\mathcal{F}} \hat{L}_n(f)$, which minimizes the empirical error, is affected by a generalization bias $(L(f^*) - \hat{L}_n(f^*))$. However, the generalization bias of a classifier can be studied by considering its supremum respect to the class of functions $\mathcal{F}$, $\sup_{f\in\mathcal{F}}[L(f) - \hat{L}_n(f)]$, which can be analyzed through MD or RC approaches [5]. The first one can be computed by shuffling and splitting the dataset in two halves:

$$\hat{\mathrm{MD}}(\mathcal{F}) = \max_{f\in\mathcal{F}} \left(\hat{L}^{(1)}_{\frac{n}{2}}(f) - \hat{L}^{(2)}_{\frac{n}{2}}(f)\right) \tag{1}$$

where $\hat{L}^{(1)}_{\frac{n}{2}}(f) = \frac{2}{n}\sum_{i=1}^{\frac{n}{2}} \ell\left(f(\boldsymbol{x}_i), y_i\right)$ and $\hat{L}^{(2)}_{\frac{n}{2}}(f) = \frac{2}{n}\sum_{i=\frac{n}{2}+1}^{n} \ell\left(f(\boldsymbol{x}_i), y_i\right)$. Alternatively, RC [6] can be computed from the training set by randomly reassigning the labels of the patterns:

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

where $\sigma_i \in \{-1, +1\}$ with $\mathcal{P}(\sigma_i = +1) = \mathcal{P}(\sigma_i = -1) = 1/2$. Based on the previous quantities, it is then possible to prove the two following bounds for $L(f)$ [5], which hold with probability $(1-\delta)$:

$$L(f) \ \leq \ L^{MD}(f) = \hat{L}_n(f) + \frac{1}{m}\sum_{i=1}^{m} \hat{\mathrm{MD}}^{(i)}(\mathcal{F}) + 3\sqrt{\frac{\log\frac{2}{\delta}}{2n}} \tag{3}$$

$$L(f) \ \leq \ L^{RC}(f) = \hat{L}_n(f) + \hat{\mathrm{RC}}(\mathcal{F}) + 3\sqrt{\frac{\log\frac{2}{\delta}}{2n}}. \tag{4}$$

Note that, in our formulation, the value of Eq. (3) is computed by repeating the procedure m times, $m \leq \binom{n}{n/2}$, so to avoid possible "unlucky" splittings [8].

3 Maximal Discrepancy Vs. Rademacher Complexity

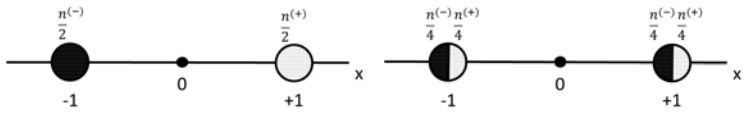
To the best knowledge of the authors, it was never established if the MD outperforms the RC one or vice versa. In other words, it is not known which approach produces the tightest bound. We will show that the two methods complement

ESANN 2011 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 27-29 April 2011, i6doc.com publ., ISBN 978-2-87419-044-5.
Available from http://www.i6doc.com/en/livre/?GCOI=28001100817300.

each other, in the sense that they provide different results, depending on the difficulty of the training problem.

In order to better understand their behavior, we build two artificial problems that represent two extreme cases: the first one is a trivial linearly separable problem, while the second one consists of two completely overlapped classes. For simplicity, the artificial problem makes use of mono-dimensional datasets: the results, as described in the following sections, are confirmed with high-dimensional datasets as well. All the samples are centered in ± 1: the probability function generating the data is such that $\mathcal{P}(x = +1) = \mathcal{P}(x = -1) = 1/2$ and $\mathcal{P}(x \neq \pm 1) = 0$. The two artificial sets X_{a1}, X_{a2}, are depicted in Fig. 1:

1. the patterns of X_{a1} are such that $(x_i, y_i)_{a1} = (+1, +1)$ if $i \in [1, n/2]$, and $(x_i, y_i)_{a1} = (-1, -1)$ otherwise;
2. the patterns of X_{a2} are such that: $(x_i, y_i)_{a2} = (+1, +1)$ if $i \in [1, n/4]$, $(x_i, y_i)_{a2} = (+1, -1)$ if $i \in [n/4 + 1, n/2]$, $(x_i, y_i)_{a2} = (-1, +1)$ if $i \in n/2 + 1, 3n/4]$, and $(x_i, y_i)_{a2} = (-1, -1)$ if $i \in [3n/4 + 1, n]$.



(a) X_{a1} (b) X_{a2}

Fig. 1: The artificial datasets used for comparing MD and RC.

We consider the *hard loss function* $\ell_H\left(f(\boldsymbol{x}_i), y_i\right) = \mathbf{1}\left\{y_i f(\boldsymbol{x}_i)\right\}$, which exploits the *indicator function* $\mathbf{1}(\cdot, \cdot)$, so that the optimal classifier f^* is selected according to the empirical error. In fact we can take into account only four possible classifiers: (i) $f(x) = +1$; (ii) $f(x) = +x$; (iii) $f(x) = -x$; and (iv) $f(x) = -1$. By considering all the possible $\binom{n}{n/2}$ shufflings in Eq. (3) and all the possible 2^n combinations of labels in Eq. (4), we can precisely compare the MD-based and RC-based bounds.

Table 1 shows the value of the empirical error, which also represents the best misclassification rate for the datasets (i.e. $L(f) = \hat{L}_n(f^*)$), and the error estimations $L^{MD}(f)$ and $L^{RC}(f)$, computed using Eqns. (3) and (4), respectively. The term depending on δ is omitted, as it is constant, once n is fixed. The value of $\hat{\mathrm{RC}}(\mathcal{F})$ does not depend on the distribution $\mathcal{P}(y|x)$, as predicted by theory (Eq. (2)): thus, the same error is obtained for the two artificial sets and the estimation $L^{RC}(f)$ outperforms the MD-based bound in the case of highly overlapped classes (i.e. on X_{a2}). On the contrary, the performance of the MD-based error estimation is noticeably better than $L^{RC}(f)$ when the two classes are linearly separable (X_{a1}), as $L^{MD}(f)$ takes into account the characteristics of the unknown $\mathcal{P}(x, y)$.

Both approaches provide loose estimations, even on these simple artificial problems. However, we propose a method to tighten the MD- and RC-based

ESANN 2011 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 27-29 April 2011, i6doc.com publ., ISBN 978-2-87419-044-5. Available from http://www.i6doc.com/en/livre/?GCOI=28001100817300.

(a) Results obtained on X_{a1}.

n	$L(f)$	$L^{MD}(f)$	$L^{RC}(f)$
10	0.0	**28.6**	37.5
20	0.0	**17.1**	24.6
30	0.0	**15.2**	21.0

(b) Results obtained on X_{a2}.

n	$L(f)$	$L^{MD}(f)$	$L^{RC}(f)$
10	50.0	89.0	**87.5**
20	50.0	75.3	**74.6**
30	50.0	71.3	**71.0**

Table 1: Error estimations with MD and RC on the two artificial datasets. All results are in percentage, best values are in bold face.

bounds, so that it is possible to use them in practical applications: the idea is to split the original dataset in two almost homogeneous subsets. In fact, as predicted by theory [5], the effect of creating two homogeneous subsets is to decrease the $\hat{\text{MD}}$ and $\hat{\text{RC}}$ terms of Eqns. (1) and (2). When the MD-based method is applied, the labels are flipped on half of the data in each subset; when the RC-based bound is computed, each subset is assigned to one class. Then $L_h^{MD}(f)$ is the new estimate, where the term $\hat{\text{MD}}_h(\mathcal{F})$ is computed using the previously described procedure; similarly, we compute $\hat{\text{RC}}_h(\mathcal{F})$ and, consequently, $L_h^{RC}(f)$. In general, any procedure which allows to divide a dataset in two homogenenous parts can be used, such as the Nearly Homogeneous Multi-Partitioning (NHMP) technique presented in [10]. The results, presented in Table 2, confirm the effectiveness of this approach: the two bounds give the same estimations and reach the true error value $L(f)$.

(a) Results obtained on X_{a1}.

n	$L(f)$	$L_h^{MD}(f)$	$L_h^{RC}(f)$
10	0.0	**0.0**	**0.0**
20	0.0	**0.0**	**0.0**
30	0.0	**0.0**	**0.0**

(b) Results obtained on X_{a2}.

n	$L(f)$	$L_h^{MD}(f)$	$L_h^{RC}(f)$
10	50.0	**50.0**	**50.0**
20	50.0	**50.0**	**50.0**
30	50.0	**50.0**	**50.0**

Table 2: Error estimations with MD_h and RC_h on the two artificial datasets. Best results are in bold face.

In conclusion, we can claim that the MD approach exploits the characteristics of the unknown distribution $\mathcal{P}(y|x)$ (see Eq. (1)), thus is characterized by the best performance when the two classes are easily separable. On the contrary, the value of the RC estimation is independent of $\mathcal{P}(y|x)$ (but depends on $\mathcal{P}(x)$, see Eq. (2)), thus provides tighter estimates in the case of highly overlapped classes. The homogenizing procedure allows to improve the estimations and to predict the true error value on the artificial datasets.

Unfortunately, it can be shown that the two methods are not sufficient to obtain tight bounds in practice [9]. However, we can exploit the error estimation as a guide for tuning additional (hyper-)parameters, required for building an optimal classifier (i.e. for model selection purposes). In particular, we consider linear classifiers, trained using the *Support Vector Machine* (SVM) [3], which requires the tuning of a hyper-parameter (C). In particular, the *peeled* version

ESANN 2011 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 27-29 April 2011, i6doc.com publ., ISBN 978-2-87419-044-5. Available from http://www.i6doc.com/en/livre/?GCOI=28001100817300.

of SVM [7, 8, 9] will be used, as it allows to rigorously compute the MD and RC bounds.

(a) MNIST

n	MD	RC
10	**2.5** ± 0.6	2.6 ± 0.6
20	**2.4** ± 0.3	2.5 ± 0.3
40	**1.2** ± 0.3	1.3 ± 0.3
60	**0.6** ± 0.1	0.9 ± 0.2
80	**0.7** ± 0.2	0.8 ± 0.2
100	**0.5** ± 0.1	0.6 ± 0.1

(b) DaimlerChrysler

n	MD	RC
10	24.9 ± 0.9	**24.8** ± 0.9
20	29.8 ± 0.7	**29.0** ± 0.7
40	27.2 ± 0.9	**26.1** ± 1.0
60	27.3 ± 0.8	**25.9** ± 0.7
80	25.3 ± 0.8	**24.9** ± 0.8
100	25.7 ± 0.5	**24.6** ± 0.6

Table 3: Test error rates obtained using MD and RC on real–world datasets. All results are in percentage, best values are in bold face.

(a) MNIST

n	MD_h	RC_h
10	**2.3** ± 0.5	2.5 ± 0.5
20	**1.4** ± 0.2	**1.4** ± 0.2
40	**0.5** ± 0.1	0.6 ± 0.1
60	**0.4** ± 0.1	0.5 ± 0.1
80	**0.4** ± 0.1	0.5 ± 0.1
100	**0.4** ± 0.1	**0.4** ± 0.1

(b) DaimlerChrysler

n	MD_h	RC_h
10	**28.6** ± 1.5	**28.6** ± 1.5
20	**29.5** ± 0.9	**29.5** ± 0.9
40	**22.2** ± 0.6	**22.2** ± 0.6
60	**21.4** ± 0.5	21.5 ± 0.5
80	**20.6** ± 0.3	**20.6** ± 0.3
100	20.7 ± 0.4	**20.6** ± 0.4

Table 4: Test error rates obtained using MD_h and RC_h on real–world datasets, after applying the NHMP procedure. All results are in percentage, best values are in bold face.

4 Error estimation for model selection

We consider the MNIST [11] and the DaimlerChrysler [12] datasets, consisting of a large number of samples, and use only a small amount of patterns for training purposes, so that the test error rate computed on the remaining patterns is a good approximation of the true error $L(f)$. Concerning MNIST, we consider the 13074 patterns containing 0's and 1's, allowing us to deal with a binary classification problem. The DaimlerChrysler dataset consists of 9800 grayscale images, representing pedestrians crossing a road and non-pedestrian examples. While the MNIST dataset is known to be an almost linearly separable problem, the two classes of the DaimlerChrysler dataset are highly overlapped: therefore, these two problems are the real–world counterparts of X_{a1} and X_{a2}.

Tables 3(a) and 3(b) show the average test error rates obtained by performing the model selection, according to the MD and RC error estimations: $m = 40$ is used for the MD term of Eq. (3), while the expectation in Eq. (2) is computed

ESANN 2011 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 27-29 April 2011, i6doc.com publ., ISBN 978-2-87419-044-5. Available from http://www.i6doc.com/en/livre/?GCOI=28001100817300.

through a Monte Carlo procedure (100 trials). The results confirm that MD outperforms RC when the two classes are linearly separable (i.e. MNIST), while the opposite is true when the two classes overlap (i.e. DaimlerChrysler). Tables 4(a) and 4(b) show the effect of the homogenizing procedure: in this case, like in the artificial one of previous Section, the two methods perform similarly. The main advantage of this approach lies in its ability to identify a better performing classifier, as shown by comparing these results with the corresponding ones in Tables 3(a) and 3(b). More details can be found in [9].

5 Concluding remarks

In this paper we have shown that there is a complementarity between the Maximal Discrepancy and the Rademacher Complexity, when estimating the generalization error of a classifier but, in general, no method outperforms the other. In fact, MD behaves better when applied to easy separable problems, while RC obtain tighter bounds on difficult ones. From a practical point of view, both methods are effective in model selection and the use of a homogenizing procedure to the dataset allows to further improve their performance.

References

[1] A. Blum, A. Kalai, and J. Langford. Beating the hold–out: Bounds for k–fold and progressive cross–validation. In *Proc. of the Conference on Learning Theory*, pages 203–208, 1999.

[2] A. Isaksson, M. Wallman, H. Goeransson, and M.G. Gustafsson. Cross–validation and bootstrapping are unreliable in small sample classification. *Pattern Recognition Letters*, 29:1960–1965, 2008.

[3] V.N. Vapnik. *The Nature of Statistical Learning Theory*. Springer, 2000.

[4] T. Poggio, R. Rifkin, S. Mukherjee, and P. Niyogi. General conditions for predictivity in learning theory. *Nature*, 428:419–422, 2004.

[5] P.L. Bartlett, S. Boucheron, and G. Lugosi. Model selection and error estimation. *Machine Learning*, 48:85–113, 2002.

[6] P. Bartlett and S. Mendelson. Rademacher and Gaussian complexities: Risk bounds and structural results. *Journal of Machine Learning Research*, 3:463–482, 2002.

[7] D. Anguita, A. Ghio, N. Greco, L. Oneto, and S. Ridella. Model selection for support vector machines: Advantages and disadvantages of the machine learning theory. In *Proc. of the Int. Joint Conference on Neural Networks 2010*, 2010.

[8] D. Anguita, A. Ghio, and S. Ridella. Maximal Discrepancy for Support Vector Machines. In *Proc. of European Symposium on Artificial Neural Networks 2010*, 2010.

[9] D. Anguita, A. Ghio, L. Oneto, and S. Ridella. Maximal Discrepancy Vs. Rademacher Complexity for Error Estimation. Technical report, University of Genoa - available for download at http://smartlab.dibe.unige.it/publications_tr.aspx, 2010.

[10] M. Aupetit. Nearly homogeneous multi-partitioning with a deterministic generator. *Neurocomputing*, 72:1379–1389, 2009.

[11] H. Larochelle, D. Erhan, A. Courville, J. Bergstra, and Y. Bengio. An empirical evaluation of deep architectures on problems with many factors of variation. In *Proc. of the International Conference on Machine Learning*, pages 473–480, 2007.

[12] S. Munder and D.M. Gavrila. An Experimental Study on Pedestrian Classification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28:1863–1868, 2006.