ESANN 2018 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 25-27 April 2018, i6doc.com publ., ISBN 978-287587047-6. Available from http://www.i6doc.com/en/.

A Sharper Bound on the Rademacher Complexity of Margin Multi-category Classifiers

Khadija Musayeva, Fabien Lauer and Yann Guermeur

LORIA, University of Lorraine, CNRS
Nancy, France
{khadija.musayeva,fabien.lauer,yann.guermeur}@loria.fr

Abstract. One of the main open problems in the theory of margin multi-category pattern classification is the dependency of a guaranteed risk on the number C of categories, the sample size m and the margin parameter γ. This paper derives a new bound on the probability of error of margin multi-category classifiers under minimal learnability assumptions. It improves the dependency on C over the state of the art. This is achieved through the introduction of a new Sauer-Shelah lemma.

1 Introduction

One of the main open problems in the theory of margin multi-category pattern classification is the dependency of a guaranteed risk on the number C of categories, the sample size m and the margin parameter γ. In this paper, we focus on the dependency on the first parameter when minimal learnability assumptions are made. One of the approaches to bound the risk of margin multi-category classifiers, especially efficient in obtaining data-dependent bounds, starts with a basic supremum inequality involving the Rademacher complexity [1]. The use of this pathway can also be justified by the availability of a rich toolset from the theory of Gaussian processes, as demonstrated in [2]. Using a structural result for the Rademacher complexity, a linear dependency on C was obtained in [3], improving upon the bound of [1]. Yet, as shown in [4], linking the Rademacher complexity to metric entropies by the chaining method [5] and postponing the decomposition to this level, opens up the possibility to obtain bounds sublinear in C. Here, we precisely follow the pathway of [4]. In this context, our contribution is the introduction of a new metric entropy bound generalizing that of [6]. This leads to an improved dependency on the number of categories over that of [4]. More precisely, we exchange a power of C by a power of $\ln(C)$ while maintaining the same dependency on m and γ.

Formally, we consider C-category pattern classification problems with $C \geqslant 3$. We denote by $[\![i, j]\!]$ the set of integers from i to j. Each object is represented by its description $x \in \mathcal{X}$ and the categories y belong to $\mathcal{Y} = [\![1, C]\!]$. We assume that the link between descriptions and categories can be characterized by an unknown probability measure P on $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$. Let $Z = (X, Y)$ be a random pair with values in $\mathcal{Z}$, distributed according to P. The available information on P is limited to an m-sample $\mathbf{Z}_m = (Z_i)_{1 \leqslant i \leqslant m} = ((X_i, Y_i))_{1 \leqslant i \leqslant m}$ distributed according to P^m and we make the hypothesis that $m > C$. In the following, we

ESANN 2018 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 25-27 April 2018, i6doc.com publ., ISBN 978-287587047-6. Available from http://www.i6doc.com/en/.

distinguish the sample size m from the generic notation n which stands for a number of points in a set that needs not be a realization of a random sample.

We consider margin classifiers that take their decisions based on a score per category and focus on those that implement classes of functions with values in a hypercube of $\mathbb{R}^C$.

Definition 1 (Margin multi-category classifiers). *Let $\mathcal{G} = \prod_{k=1}^{C} \mathcal{G}_k$ be a class of functions from $\mathcal{X}$ into $[-M_{\mathcal{G}}, M_{\mathcal{G}}]^C$ with $M_{\mathcal{G}} \in [1, +\infty)$. For each function $g = (g_k)_{1 \leqslant k \leqslant C} \in \mathcal{G}$ and $x \in \mathcal{X}$, a* margin multi-category classifier *outputs* $\text{argmax}_{1 \leqslant k \leqslant C}\, g_k(x)$.

The basic supremum inequality mentioned above involves the Rademacher complexity of a class of margin functions built upon $\mathcal{G}$. We use a variant of the one used in [3], discarding all information irrelevant to the characterization of classification accuracy (the values above the margin parameter γ, as well as the ones below zero). The use of this version results in a tighter bound.

Definition 2 (Class of functions $\mathcal{F}_{\mathcal{G},\gamma}$). *Let $\mathcal{G}$ be a class of functions satisfying Definition 1. For every $\gamma \in (0, 1]$, the class $\mathcal{F}_{\mathcal{G},\gamma}$ is*

$$\left\{ f_{g,\gamma} \in [0, \gamma]^{\mathcal{Z}} : f_{g,\gamma}(x, k) = \max\left(0, \min\left(\gamma, \frac{1}{2}(g_k(x) - \max_{l \neq k} g_l(x))\right)\right), g \in \mathcal{G} \right\}.$$

Hereafter, $\mathcal{F}$ is a class of real-valued functions on a measurable space $\mathcal{T}$. Now, recall the definition of the Rademacher complexity. Let $\mathbf{T}_n$ be a sequence $(T_i)_{1 \leqslant i \leqslant n}$ of i.i.d. random variables taking their values in $\mathcal{T}$ and $\boldsymbol{\sigma}_n$ a sequence $(\sigma_i)_{1 \leqslant i \leqslant n}$ of i.i.d. random variables uniformly distributed in $\{-1, 1\}$. Then, the empirical Rademacher complexity of $\mathcal{F}$ given $\mathbf{T}_n$ is defined as

$$\hat{R}_n(\mathcal{F}) = \mathbb{E}_{\boldsymbol{\sigma}_n}\left[\sup_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} \sigma_i f(T_i) \,\middle|\, \mathbf{T}_n \right]$$

and its Rademacher complexity is $R_n(\mathcal{F}) = \mathbb{E}_{\mathbf{T}_n}\left[\hat{R}_n(\mathcal{F})\right]$.

Another capacity measure appearing in our bounds is the fat-shattering dimension also known as the γ-dimension. It is defined as follows. For $\gamma \in \mathbb{R}_+^*$, a subset $s_{\mathcal{T}^n} = \{t_i : 1 \leqslant i \leqslant n\}$ of $\mathcal{T}$ is said to be *γ-shattered* by $\mathcal{F}$ if there is a vector $\mathbf{b}_n = (b_i)_{1 \leqslant i \leqslant n} \in \mathbb{R}^n$ such that, for every vector $\mathbf{s}_n = (s_i)_{1 \leqslant i \leqslant n} \in \{-1, 1\}^n$, there is a function $f_{\mathbf{s}_n} \in \mathcal{F}$ satisfying: $\forall i \in [\![1, n]\!]$, $\;s_i(f_{\mathbf{s}_n}(t_i) - b_i) \geqslant \gamma$. The *fat-shattering dimension with margin γ* of the class $\mathcal{F}$, $\gamma\text{-dim}(\mathcal{F})$, is the maximal cardinality of a subset of $\mathcal{T}$ γ-shattered by $\mathcal{F}$, if such maximum exists. Otherwise it is infinite. As in [4, 7], we make the following hypothesis regarding the fat-shattering dimensions.

Hypothesis 1. *We consider classes of functions $\mathcal{G}$ satisfying Definition 1 plus the fact that there exists a pair $(d_{\mathcal{G}}, K_{\mathcal{G}}) \in \left(\mathbb{R}_+^*\right)^2$ such that*

$$\forall \epsilon \in (0, M_{\mathcal{G}}], \quad \max_{1 \leqslant k \leqslant C} \epsilon\text{-}dim(\mathcal{G}_k) \leqslant K_{\mathcal{G}} \epsilon^{-d_{\mathcal{G}}}.$$

ESANN 2018 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 25-27 April 2018, i6doc.com publ., ISBN 978-287587047-6. Available from http://www.i6doc.com/en/.

The capacity measures that connect Rademacher complexities with fat-shattering dimensions are covering numbers. For any $f, f' \in \mathcal{F}$ and $\mathbf{t}_n = (t_i)_{1 \leqslant i \leqslant n} \in \mathcal{T}^n$, let

$$d_{p,\mathbf{t}_n}(f, f') = \begin{cases} \left(\frac{1}{n}\sum_{i=1}^{n} |f(t_i) - f'(t_i)|^p\right)^{\frac{1}{p}}, \text{if } p \in [1, +\infty) \\ \max_{1 \leqslant i \leqslant n} |f(t_i) - f'(t_i)|, \text{if } p = +\infty. \end{cases}$$

Then, the covering number of $\mathcal{F}$ at scale $\epsilon > 0$ with respect to $d_{p,\mathbf{t}_n}$, $\mathcal{N}(\epsilon, \mathcal{F}, d_{p,\mathbf{t}_n})$, is the smallest cardinality of ϵ-nets of $\mathcal{F}$, i.e., subsets $\bar{\mathcal{F}} \subseteq \mathcal{F}$ such that $\forall f \in \mathcal{F}$, $d_{p,\mathbf{t}_n}(f, \bar{\mathcal{F}}) < \epsilon$. The *metric entropy* of $\mathcal{F}$ is the binary logarithm of its covering number. The distribution-free nature of metric entropy bounds calls for the use of uniform covering numbers defined as $\mathcal{N}_p(\epsilon, \mathcal{F}, n) = \sup_{\mathbf{t}_n \in \mathcal{T}^n} \mathcal{N}(\epsilon, \mathcal{F}, d_{p,\mathbf{t}_n})$.

The derivation of our bound is based on the following transitions between the aforementioned capacity measures. We relate the empirical Rademacher complexity of $\mathcal{F}_{\mathcal{G},\gamma}$ to its metric entropy through the chaining method [5] as

$$\hat{R}_n(\mathcal{F}_{\mathcal{G},\gamma}) \leqslant h(N) + 2\sum_{j=1}^{N}(h(j) + h(j-1))\sqrt{\frac{\ln \mathcal{N}(h(j), \mathcal{F}_{\mathcal{G},\gamma}, d_{2,\mathbf{z}_n})}{n}}, \quad (1)$$

where $N \in \mathbb{N}^*$ and $h : \mathbb{N}^* \to \mathbb{R}_+^*$ is a decreasing function such that $h(0)$ is greater than the diameter of $\mathcal{F}_{\mathcal{G},\gamma}$ with respect to $d_{2,\mathbf{z}_n}$. The metric entropy of $\mathcal{F}_{\mathcal{G},\gamma}$ is then related to the ones of the component function classes $\mathcal{G}_k$ by the *decomposition lemma* (Lemma 1 in [4]):

$$\forall p \in [1, +\infty], \ \ln \mathcal{N}(\epsilon, \mathcal{F}_{\mathcal{G},\gamma}, d_{p,\mathbf{z}_n}) \leqslant \sum_{k=1}^{C} \ln \mathcal{N}\left(\frac{\epsilon}{C^{1/p}}, \mathcal{G}_k, d_{p,\mathbf{x}_n}\right). \quad (2)$$

Finally, a Sauer-Shelah lemma upper bounds the metric entropies of the classes $\mathcal{G}_k$ in terms of their fat-shattering dimensions.

Using (2) with $p = 2$ and the Sauer-Shelah lemma in L_2-norm (Theorem 1 of [6]) in the chaining results in a sublinear dependency on C (Theorem 7 in [4]). On the other hand, from (2) it is clear that the dependency on C in the scale of the covering numbers disappears when one resorts to the extreme case, that is, $p = \infty$. Then, using Lemma 3.5 of [8] in the chaining (based on the straightforward relationship between the norms), a radical dependency on C can be obtained irrespective of the value of $d_{\mathcal{G}}$. On the downside, due to the fact that this metric entropy bound involves $\ln^2(\epsilon^{-1})$, the convergence rate is worsened compared to the one obtained with an L_2-norm bound involving $\ln(\epsilon^{-1})$. In the sequel, we generalize the metric entropy bound of [6] to L_p-norms with integer $p \in (2, \infty)$. We show that this generalization can be used in the chaining in combination with (2) to yield an improved dependency on C compared to Theorem 7 of [4] (without worsening the convergence rate nor the dependency on γ).

2 L_p-norm Sauer-Shelah lemma

Our generalization of Theorem 1 of [6] to $p \in (2, \infty)$ is the following one.

ESANN 2018 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 25-27 April 2018, i6doc.com publ., ISBN 978-287587047-6. Available from http://www.i6doc.com/en/.

Lemma 1. *Let $\mathcal{F}$ be a class of functions from $\mathcal{T}$ into $[-M_{\mathcal{F}}, M_{\mathcal{F}}]$ with $M_{\mathcal{F}} \in [1, +\infty)$. For $\epsilon \in (0, M_{\mathcal{F}}]$, let $d(\epsilon) = \epsilon$-dim$(\mathcal{F})$. For any integer $p > 2$, suppose that $\epsilon \in (0, 2M_{\mathcal{F}}]$ is such that $d\left(\frac{\epsilon}{M_{\mathcal{F}}+26p^2}\right)$ is finite. Then,*

$$\ln \mathcal{N}_p(\epsilon, \mathcal{F}, n) \leqslant 10p\, d\left(\frac{\epsilon}{M_{\mathcal{F}} + 26p^2}\right) \ln\left(\frac{8p^{\frac{2}{7}} M_{\mathcal{F}}^2}{\epsilon}\right).$$

Proof sketch. The proof is essentially that of Theorem 1 in [6], with the following main changes. First, we replace Lemma 4 of [6] by a consequence of Minkowski's inequality which states that for any $p \in (2, \infty)$, for any random variable T and its independent copy T',

$$\left(\mathbb{E}|T - T'|^p\right)^{1/p} \leq \left(\mathbb{E}|T|^p\right)^{1/p} + \left(\mathbb{E}|-T'|^p\right)^{1/p} = 2\left(\mathbb{E}|T|^p\right)^{1/p}.$$

With this change at hand, the computation of the integral involved in Eq. (4) of [6] produces the quantity $K_p = \sum_{k\geqslant 1} k^p/2^k$. Second, we extend the construction of a separating tree to L_p-norm. This leads to a change in the separation of trees now involving $K_p^{1/p}$. Third, for the probabilistic extraction, we make use of an L_p-norm extension of Lemma 13 in [6]: Lemma 8 in [4]. To complete the proof, we show that, since p is an integer, $K_p < p^{2p}$. To this end, note that K_p is a polylogarithm of negative order: $K_p = Li_{-p}(1/2)$. According to Lemma 1 in [9], the latter can be expressed using Stirling's numbers of the second kind as $\sum_{k=0}^{p} k! \left\{ \begin{array}{c} p+1 \\ k+1 \end{array} \right\}$. Then Theorem 3 in [10] and the fact that for $p > 2$, $(p+1) < 3p/2$ and $p^{p-1} = p^{2p}/p^{p+1} < p^{2p}/4$, give the claimed bound on K_p. $\square$

From (2) one can see that, based on $C^{\frac{1}{p}} = 2^{\left(\frac{1}{p}\log_2(C)\right)}$, the dependency on C in the scale parameter can be removed for all $p \geqslant \log_2(C)$. Now, using $p = \lceil \log_2(C) \rceil$ for $C > 4$, we obtain the following bound.

Lemma 2. *Let $\mathcal{G}$ be a class of functions satisfying Definition 1. For $\gamma \in (0, 1]$, let $\mathcal{F}_{\mathcal{G},\gamma}$ be the class of functions deduced from $\mathcal{G}$ according to Definition 2. For $\epsilon \in (0, M_{\mathcal{G}}]$, let $d(\epsilon) = \max_{1\leqslant k\leqslant C} \epsilon$-dim$(\mathcal{G}_k)$. Then, for $\epsilon \in (0, \gamma]$ and $C > 4$,*

$$\ln \mathcal{N}_p(\epsilon, \mathcal{F}_{\mathcal{G},\gamma}, m) \leqslant 10C \log_2(2C)\, d\left(\frac{\epsilon}{2M_{\mathcal{G}} + 52\log_2^2(2C)}\right) \ln\left(\frac{16 \log_2^{\frac{2}{7}}(2C)\, M_{\mathcal{G}}^2}{\epsilon}\right).$$

Proof. The claim follows from the application of (2) and Lemma 1 together with the choice $p = \lceil \log_2(C) \rceil$ and the fact that $C^{1/\lceil \log_2(C) \rceil} < 2$ and $\lceil \log_2(C) \rceil < \log_2(2C)$. $\square$

Lemma 1 provides a metric entropy bound in $O\left(d(\epsilon) \ln\left(\epsilon^{-1}\right)\right)$ as $\epsilon \to 0$, an improvement over Lemma 2 of [4] and Lemma 3.5 of [8] (see the problem pointed out at the end of Section 1). In addition, the formula of Lemma 2 exhibits a better dependency on C than the one obtained in [4]. As demonstrated below, these improvements allow us to obtain a better bound on the Rademacher complexity of $\mathcal{F}_{\mathcal{G},\gamma}$.

ESANN 2018 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 25-27 April 2018, i6doc.com publ., ISBN 978-287587047-6. Available from http://www.i6doc.com/en/.

3 Improved dependency on the number of categories

Applying our new metric entropy bound along with Hypothesis 1 in the chaining yields the following result.

Theorem 1. *Let $\mathcal{G}$ be as in Definition 1 and, for any $\gamma \in (0,1]$, $\mathcal{F}_{\mathcal{G},\gamma}$ be deduced from $\mathcal{G}$ as in Definition 2. Then, under Hypothesis 1, there is a function $K(d_{\mathcal{G}},\gamma)$ such that for all $C > 4$,*

$$R_m(\mathcal{F}_{\mathcal{G},\gamma}) \leqslant K(d_{\mathcal{G}},\gamma)\sqrt{\frac{C}{m}} \begin{cases} (\ln(C))^{d_{\mathcal{G}}+\frac{1}{2}}, & \text{if } 0 < d_{\mathcal{G}} < 2 \\ \ln^3(C)\ln^{\frac{3}{2}}\left(\frac{m}{C}\right), & \text{if } d_{\mathcal{G}} = 2 \\ m^{\frac{1}{2}-\frac{1}{d_{\mathcal{G}}}}\ln^3(C)\ln^{\frac{1}{2}}\left(\frac{m^{\frac{1}{d_{\mathcal{G}}}}}{\ln(C)^{\frac{1}{d_{\mathcal{G}}}}}\right), & \text{if } d_{\mathcal{G}} > 2. \end{cases}$$

Proof sketch. The proof closely follows that of Theorem 7 in [4]. Applying the formula of Lemma 2 and Hypothesis 1 in the chaining formula (1) yields

$$\begin{aligned} \hat{R}_m(\mathcal{F}_{\mathcal{G},\gamma}) &\leqslant h(N) + 2\sqrt{\frac{10C\log_2(2C)}{m}} \\ &\cdot \sum_{j\in\mathcal{J}}(h(j)+h(j-1))\left[d\left(\frac{h(j)}{2M_{\mathcal{G}}+52\log_2^2(2C)}\right)\ln\left(\frac{16M_{\mathcal{G}}^2\log_2^{\frac{2}{7}}(2C)}{h(j)}\right)\right]^{1/2} \\ &\leqslant h(N) + 2\sqrt{\frac{10C\log_2(2C)K_{\mathcal{G}}}{m}}\left(2M_{\mathcal{G}}+52\log_2^2(2C)\right)^{\frac{d_{\mathcal{G}}}{2}} \\ &\cdot \sum_{j\in\mathcal{J}}\frac{(h(j)+h(j-1))}{(h(j))^{\frac{d_{\mathcal{G}}}{2}}}\ln^{\frac{1}{2}}\left(\frac{16M_{\mathcal{G}}^2\log_2^{\frac{2}{7}}(2C)}{h(j)}\right), \end{aligned} \tag{3}$$

where $\mathcal{J} = \{j \in [\![1,N]\!] : h(j) \leqslant \gamma\}$. Now, depending on the value of $d_{\mathcal{G}}$, we choose N and the function h in such a way so as to optimize the dependency on C and m. When $d_{\mathcal{G}} < 2$, (3) is upper bounded by an integral and we perform exactly the same computations as in [4]. For $d_{\mathcal{G}} \geqslant 2$, we use similar computations but with a different setting than that in [4]. Namely,

$$\begin{cases} h(j)=\gamma 2^{(N-j)}\log_2^{\frac{2}{7}}(2C)\sqrt{\frac{C}{m}} \text{ and } N=\left\lceil\log_2\sqrt{\frac{m}{C}}\right\rceil, \text{if } d_{\mathcal{G}}=2 \\ h(j)=\gamma 2^{\frac{2(N-j)}{d_{\mathcal{G}}-2}}\frac{\log_2^2(2C)^{\frac{1}{d_{\mathcal{G}}}}}{m^{\frac{1}{d_{\mathcal{G}}}}} \text{ and } N=\left\lceil\frac{d_{\mathcal{G}}-2}{2d_{\mathcal{G}}}\log_2\left(\frac{m}{\log_2^{2d_{\mathcal{G}}}(2C)^{\frac{1}{d_{\mathcal{G}}}}}\right)\right\rceil, \text{otherwise.} \end{cases}$$

□

In comparison with Theorem 7 of [4], Theorem 1 replaces a power of C by a power of its logarithm. That is, $C^{\frac{1}{4}}$ is replaced by $\ln(C)$ for $d_{\mathcal{G}} < 2$ and $\sqrt{C}$ by $\ln^3(C)$ for $d_{\mathcal{G}} = 2$. For the final case, the comparison of the two bounds is less straightforward, since the term $C^{\frac{1}{d_{\mathcal{G}}}}\ln^{\frac{1}{2}}(m/C)$ is replaced by $\ln^3(C)\ln^{\frac{1}{2}}\left(m^{\frac{1}{d_{\mathcal{G}}}}/\ln(C)^{\frac{1}{d_{\mathcal{G}}}}\right)$.

4 Conclusion and future work

We derived a sharper bound on the Rademacher complexity of margin multi-category classifiers under minimal learnability assumptions. Central to this is the generalization of the metric entropy bound of [6] to L_p-norms with integer $p \in (2, +\infty)$. When applied in the chaining combined with the decomposition for metric entropies, it results in an improved dependency on C compared to [4], without worsening the convergence rate nor the dependency on the margin parameter γ. Following a similar pathway, future work will focus on obtaining bounds on the Rademacher complexity of specific sets of classifiers, such as multi-class support vector machines. The conjecture is that tighter bounds should result from bounding directly the covering numbers of the classes of functions of interest, i.e., without resorting to a generalized Sauer-Shelah lemma.

Acknowledgments

This work was partly funded by a CNRS research grant.

References

[1] V. Koltchinskii and D. Panchenko. Empirical margin distributions and bounding the generalization error of combined classifiers. *The Annals of Statistics*, 30(1):1–50, 2002.

[2] Y. Lei, Ü. Doğan, A. Binder, and M. Kloft. Multi-class SVMs: From tighter data-dependent generalization bounds to novel algorithms. In *NIPS 28*, pages 2026–2034, 2015.

[3] V. Kuznetsov, M. Mohri, and U. Syed. Multi-class deep boosting. In *NIPS 27*, pages 2501–2509, 2014.

[4] Y. Guermeur. L_p-norm Sauer–Shelah lemma for margin multi-category classifiers. *Journal of Computer and System Sciences*, 89:450–473, 2017.

[5] M. Talagrand. *Upper and Lower Bounds for Stochastic Processes: Modern Methods and Classical Problems*. Springer-Verlag, Berlin Heidelberg, 2014.

[6] S. Mendelson and R. Vershynin. Entropy and the combinatorial dimension. *Inventiones mathematicae*, 152:37–55, 2003.

[7] S. Mendelson. Rademacher averages and phase transitions in Glivenko-Cantelli classes. *IEEE Transactions on Information Theory*, 48(1):251–263, 2002.

[8] N. Alon, S. Ben-David, N. Cesa-Bianchi, and D. Haussler. Scale-sensitive dimensions, uniform convergence, and learnability. *Journal of the ACM*, 44(4):615–631, 1997.

[9] H.H. Pitters. On the number of segregating sites. *arXiv preprint:1708.05634*, 2017.

[10] B.C. Rennie and A.J. Dobson. On Stirling numbers of the second kind. *Journal of Combinatorial Theory*, 7(2):116–121, 1969.