A Note on Extending Generalization Bounds for Binary Large-margin Classifiers to Multiple Classes

Ürün Dogan[1], Tobias Glasmachers[2], and Christian Igel[3]

[1] Institut für Mathematik, Universität Potsdam, Germany
doganudb@math.uni-potsdam.de
[2] Institut für Neuroinformatik, Ruhr-Universität Bochum, Germany
tobias.glasmachers@ini.ruhr-uni-bochum.de
[3] Department of Computer Science, University of Copenhagen, Denmark
igel@diku.dk

Abstract. A generic way to extend generalization bounds for binary large-margin classifiers to large-margin multi-category classifiers is presented. The simple proceeding leads to surprisingly tight bounds showing the same $\tilde{O}(d^2)$ scaling in the number d of classes as state-of-the-art results. The approach is exemplified by extending a textbook bound based on Rademacher complexity, which leads to a multi-class bound depending on the sum of the margin violations of the classifier.

1 Introduction

The generalization performance of binary (two-class) large-margin classifiers is well analysed (e.g., [1–8]). The theory of generalization bounds for multi-class support vector machines (multi-class SVMs) follows the route already paved by the analysis of the binary case, for instance, in the work of Guermeur [9, 10].

In this note, we link the analysis of binary and multi-class large margin classifiers explicitly. A straightforward technique to generalize bounds for binary learning machines to the multi-class case is presented, which is based on a simple union bound argument. The next section introduces the classification framework and extensions of large-margin separation to multiple classes. Section 3 proves our main result showing how to derive bounds for multi-category classification based on bounds for the binary case. In Section 4 we apply the proposed method to a textbook result based on Rademacher complexity. The newly derived bound is discussed with respect to different multi-class SVM formulations.

2 Large-margin Multi-category Classification

We consider learning a hypothesis $h : X \to Y$ from training data

$$S = \big((x_1, y_1), \ldots, (x_\ell, y_\ell)\big) \in \big(X \times Y\big)^\ell \ ,$$

where X and Y are the input and label space, respectively. We restrict our considerations to the standard case of a finite label space (however, there exist extensions of multi-class SVMs to infinite label spaces, e.g., [11]). We denote the cardinality $|Y|$ by $d \in \mathbb{N}$. Without loss of generality we assume $Y = \{1, \ldots, d\}$ in the sequel. We presume all data points (x_n, y_n) to be sampled i.i.d. from a fixed distribution P on $X \times Y$. Then the goal of learning is to map the training data to a hypothesis h with as low as possible risk (generalization error)

$$\mathcal{R}(h) = \int_{X \times Y} \mathbf{1}_{(h(x) \neq y)} \, \mathrm{d}P(x, y) \ . \tag{1}$$

Here, the 0-1 loss is encoded by the indicator function $\mathbf{1}_{(h(x)\neq y)}$ of the set $\{(x, y) \in X \times Y \,|\, h(x) \neq y\}$.

All machines considered in this study construct hypotheses of the form

$$x \mapsto \arg\max_{c \in Y} \left[\langle w_c, \phi(x) \rangle + b_c \right] \ , \tag{2}$$

where $\phi : X \to \mathcal{H}$ is a feature map into an inner product space $\mathcal{H}$, $w_1, \ldots, w_d \in \mathcal{H}$ are class-wise weight vectors, and $b_1, \ldots, b_d \in \mathbb{R}$ are class-wise bias/offset values. The most important case is that of a feature map defined by a positive definite kernel function $k : X \times X \to \mathbb{R}$ with the property $k(x, x') = \langle \phi(x), \phi(x') \rangle$. For instance, we can set $\phi(x) = k(x, \cdot)$, in which case $\mathcal{H}$ is the corresponding reproducing kernel Hilbert space [12]. We presume that the arg max operator in equation (2) returns a single class index (ties may, e.g., be broken at random). We define the vector-valued function $f : X \to \mathbb{R}^d$ by $f = (f_1, \ldots, f_d)$ with $f_c = \langle w_c, \phi(\cdot) \rangle + b_c$ for $c \in \{1, \ldots, d\}$. Then $h(x) = \arg\max_{c \in Y} f_c(x)$ and, to ease the notation, we define $\mathcal{R}(f)$ to be equal to the corresponding $\mathcal{R}(h)$.

For a binary classifier based on thresholding a real-valued function $f : X \to \mathbb{R}$ at zero we define the hinge loss $L^{\text{hinge}}(f(x), y) = \max\{0, 1 - y \cdot f(x)\}$, a convex surrogate for the 0-1 loss used in equation (1). The expression $y \cdot f(x)$ is the (functional) margin of the training pattern (x, y). The hinge loss measures the extent to which a pattern fails to meet a target margin of one. There are different ways to extend this loss to multiple classes. Large-margin classification based on the decision function (2) can be interpreted as highlighting differences between components $f_c(x)$. This is because the difference $f_c(x) - f_e(x)$ indicates whether class c is preferred over class e in decision making. Accordingly, two canonical extensions of the hinge loss to the multi-class case $f : X \to \mathbb{R}^d$ are the sum loss

$$L^{\text{sum}}(f(x), y) = \sum_{c \in Y \setminus \{y\}} \left[L^{\text{hinge}} \left(\frac{1}{2} \big(f_y(x) - f_c(x) \big), 1 \right) \right]$$

and the maximum loss

$$L^{\max}(f(x), y) = \max_{c \in Y \setminus \{y\}} \left[L^{\text{hinge}} \left(\frac{1}{2} \big(f_y(x) - f_c(x) \big), 1 \right) \right] \ .$$

These losses are arranged so that in the binary case $d = 2$ they reduce to the hinge loss. We denote the corresponding risks by

$$\mathcal{R}^{\text{type}}(f) = \int_{X \times Y} L^{\text{type}}(f(x), y) \, \mathrm{d}P(x, y)$$

and the empirical risk for a sample $S = \big((x_1, y_1), \ldots, (x_\ell, y_\ell)\big)$ by

$$\mathcal{R}_S^{\text{type}}(f) = \frac{1}{\ell} \sum_{i=1}^{\ell} L^{\text{type}}(f(x_i), y_i) \ ,$$

where the superscript "type" is generic for "hinge", "sum", or "max".

3 Extending Bounds to Multi-category Classification

Our analysis relies on the basic insight that there are $d - 1$ distinct possible mistakes per example (x, y), namely preferring class $c \in Y \setminus \{y\}$ over the true class y. Each of these mistakes corresponds to one binary problem (having a decision function with weight vector $w_y - w_c$) indicating the specific mistake. One of these mistakes is sufficient for wrong classification, and no "binary" mistake at all implies correct classification. Then, a union bound over all mistakes gives the multi-class generalization result based on established bounds for binary classifiers. Assume that we have a bound of the following generic form:

Assumption 1 *With probability $1 - \delta$ over randomly drawn training sets S of size ℓ the risk $\mathcal{R}(f^{bin})$ of a binary classifier derived from a function $f^{bin} \in \mathbb{F}^{bin}$ is bounded by*

$$\mathcal{R}(f^{bin}) \leq B^{bin}\left(\ell, \mathcal{R}_S^{hinge}(f^{bin}), \mathcal{C}(\mathbb{F}^{bin}), \delta\right) \ ,$$

where $\mathbb{F}^{bin}$ is a space of functions $X \to \mathbb{R}$. The function $\mathcal{C}$ measures the complexity of the function class $\mathbb{F}^{bin}$ in a possibly data-dependent manner (i.e., it may implicitly depend on properties of the training data, typically in terms of the kernel Gram matrix).

Then we have:

Theorem 1 *Under Assumption 1, with probability $1 - \delta$ over randomly drawn training sets $S \in \big(X \times \{1, \ldots, d\}\big)^\ell$, the risk $\mathcal{R}(f)$ of a multi-class classifier derived from the function $f = (f_1, \ldots, f_d) : X \to \mathbb{R}^d$, $f \in \mathbb{F}$, using the decision rule (2) is bounded by*

$$\mathcal{R}(f) \leq \sum_{1 \leq c < e \leq d} \left(\frac{\ell^{(c,e)}}{\ell} + \frac{1}{\sqrt{\ell}} \sqrt{\frac{\log(d(d-1)) - \log \delta}{2}} \right) \cdot B^{bin}\left(\ell^{(c,e)}, \mathcal{R}_{S^{(c,e)}}^{hinge}\left(\frac{1}{2}(f_c - f_e)\right), \mathcal{C}(\mathbb{F}^{(c,e)}), \frac{\delta}{d(d-1)}\right) \ , \quad (3)$$

where $S^{(c,e)} = \left\{(x,y) \in S \,\middle|\, y \in \{c,e\}\right\}$ *is the training set restricted to examples of classes* c *and* e, $\ell^{(c,e)} = \left|S^{(c,e)}\right|$ *denotes its cardinality, and the pairwise binary function classes are defined as*

$$\mathbb{F}^{(c,e)} = \left\{ \frac{1}{2}(f_c - f_e) \,\middle|\, f = (f_1, \ldots, f_d) \in \mathbb{F} \right\} \quad .$$

Proof. Following the line of arguments above, the general case of d-category classification with $f = (f_1, \ldots, f_d) \in \mathbb{F}$ can be reduced to the binary case via the inequality

$$\begin{aligned} \mathcal{R}(f) &\leq \sum_{1 \leq c < e \leq d} [P(y=c) + P(y=e)] \cdot \mathcal{R}\left(\frac{1}{2}(f_c - f_e)\right) \\ &\leq \sum_{1 \leq c < e \leq d} \mathcal{R}\left(\frac{1}{2}(f_c - f_e)\right) \quad , \end{aligned}$$

where $\mathcal{R}\left(\frac{1}{2}(f_c - f_e)\right)$ refers to the risk of $\frac{1}{2}(f_c - f_e)$ solving the binary classification problem of separating class c from class e. Comparing the left to the right term of the above inequality for the risk gives the immediate result

$$\mathcal{R}(f) \leq \sum_{1 \leq c < e \leq d} B^{\mathrm{bin}}\left(\ell^{(c,e)}, \mathcal{R}^{\mathrm{hinge}}_{S^{(c,e)}}\left(\frac{1}{2}(f_c - f_e)\right), \mathcal{C}\left(\mathbb{F}^{(c,e)}\right), \frac{2\delta}{d(d-1)}\right) \quad ,$$

where we have split the probability δ over the samples into $d(d-1)/2$ equal chunks. This is conservative, since pairs of classes are highly dependent.

This bound can be refined by taking class-wise probabilities into account. By applying the Hoeffding bound we derive that $P(y=c) + P(y=e)$ is upper bounded by

$$\frac{\ell^{(c,e)}}{\ell} + \frac{1}{\sqrt{\ell}} \sqrt{\frac{\log(d(d-1)) - \log\delta}{2}}$$

with a probability of $1 - \delta'/2$ with $\delta' = 2\delta/(d(d-1))$. That is, this bound holds simultaneously for all $d(d-1)/2$ pairs of classes with a probability of $1 - \delta/2$. □

The pairwise complexity terms $\mathcal{C}(\mathbb{F}^{(c,e)})$ can be replaced with the complexity measure

$$\mathcal{C}(\mathbb{F}) = \max_{1 \leq c < e \leq d} \mathcal{C}\left(\mathbb{F}^{(c,e)}\right)$$

for $\mathbb{R}^d$-valued functions. Depending on the structure of the underlying binary bound B^{bin} the sum over all pairs of classes can be further collapsed into a factor of $d(d-1)/2$, for instance by taking the maximum over the summands.

4 Example: A Bound Based on Rademacher Complexity

Theorem 1 can be used to obtain a variety of generalization bounds when combined with the wealth of results for binary machines that can be brought in a form of Assumption 1. This section will consider an textbook generalization bound derived for binary SVMs and measuring function class flexibility by Rademacher complexity. The result will then be discussed w.r.t. two multi-class extensions of SVMs. Such a comparison can be performed with different goals in mind. On the one hand, a unifying analysis covering many different multi-class SVM types is desirable. On the other hand, one would like to see differences in the performance guarantees for different machines that may indicate superiority of one machine over another. We attempt to highlight such differences. This is in contrast to other studies such as the influential work in [9], where the goal was unification and, therefore, to make differences between machines invisible.

4.1 Extending a Binary Bound Based on Rademacher Complexity

We begin by stating the result for binary machines, in which the Rademacher complexity of real-valued functions is bounded based on the kernel Gram matrix K of the data:

Theorem 2 *Fix $\rho > 0$ and let $\mathbb{F}^{bin}$ be the class of functions in a kernel-defined feature space with norm at most $1/\rho$. Let S be a training set of size ℓ, and fix $\delta \in (0,1)$. Then with probability of at least $1-\delta$ over samples of size ℓ we have for $f^{bin} \in \mathbb{F}^{bin}$*

$$\mathcal{R}(f^{bin}) \leq B^{bin} = \mathcal{R}_S^{hinge}(f^{bin}) + \frac{4}{\ell\rho}\sqrt{\mathrm{tr}(K)} + 3\sqrt{\frac{\log(2/\delta)}{2\ell}} \ ,$$

where K is the kernel Gram matrix of the training set.

This result can be derived following the proof of Theorem 4.17 by [3]. Application of inequality (3) yields the following generalization bound:

Corollary 1 *Let $S \in \left(X \times \{1,\ldots,d\}\right)^{\ell}$ be a training set. Fix $\rho > 0$, and let $\mathbb{F}_\rho$ be the class of $\mathbb{R}^d$-valued functions in a kernel-defined feature space with semi-norm at most $1/\rho$ w.r.t. the semi-norm $\|f\| = \max\left\{\frac{1}{2}\|f_c - f_e\| \,\middle|\, 1 \leq c < e \leq d\right\}$. With probability $1-\delta$ over randomly drawn training sets $S \in \left(X \times \{1,\ldots,d\}\right)^{\ell}$, the risk $\mathcal{R}(f)$ of a multi-class classifier using the decision rule (2) is bounded by*

$$\mathcal{R}(f) \leq \sum_{1 \leq c < e \leq d} \left(\frac{\ell^{(c,e)}}{\ell} + \frac{1}{\sqrt{\ell}}\sqrt{\frac{\log(d(d-1)) - \log\delta}{2}} \right) \cdot \left[\mathcal{R}^{hinge}_{S^{(c,e)}}\left(\frac{1}{2}(f_c - f_e)\right) + \frac{4}{\ell^{(c,e)}\rho}\sqrt{\mathrm{tr}(K^{(c,e)})} + 3\sqrt{\frac{\log(2d(d-1)/\delta)}{2\ell^{(c,e)}}} \right] \ ,$$

where $K^{(c,e)}$ denotes the $\ell^{(c,e)} \times \ell^{(c,e)}$ kernel matrix restricted to examples of classes c and e.

With $\mathrm{tr}(K^{(c,e)}) \leq \mathrm{tr}(K)$ this bound reads in (not fully simplified) $\tilde{O}$-notation

$$\mathcal{R}(f) \in \tilde{O}\left(\frac{d(d-1)}{2}\left(\frac{4}{\rho \cdot \ell} \cdot \sqrt{\mathrm{tr}(K)} + \mathcal{R}_S^{\mathrm{sum}}(f)\right)\right) \ , \tag{4}$$

with the same separation of complexity and empirical risk terms as in the binary bound.[4]

4.2 Sum vs. Maximum of Margin Violations

There is no canonical extension of the binary SVM [13, 14] to multiple classes. Several slightly different formulations have been proposed, most of which reduce to the standard binary SVM if $d = 2$.

The all-in-one methods proposed independently Weston and Watkins [15], Vapnik ([1], Section 10.10), and by Bredensteiner and Bennett [16] turned out to be equivalent, up to rescaling of the decision functions and the regularization parameter $C > 0$. The method is defined by the optimization problem

$$\min \quad \frac{1}{2}\sum_{c=1}^{d} \langle w_c, w_c \rangle + C \cdot \mathcal{R}_S^{\mathrm{sum}}(f) \ .$$

An alternative multi-class SVM was proposed by Crammer and Singer [17]. It also takes all class relations into account simultaneously and solves a single optimization problem, however, penalizing the maximal margin violation instead of the sum:

$$\min \quad \frac{1}{2}\sum_{c=1}^{d} \langle w_c, w_c \rangle + C \cdot \mathcal{R}_S^{\max}(f)$$

Are there theoretical arguments to prefer one formulation over the other? In [18], the *empirical* risk of multi-class SVMs is upper bounded in terms of the empirical maximum risk $\mathcal{R}^{\max}$. This is an almost trivial result, because the hinge loss (and therefore the maximum loss) is, per construction, an upper bound on the 0-1-loss. Based on this bound it has been argued that the SVM proposed by Crammer and Singer has advantages compared to the formulation by Weston and Watkins because it leads to lower values in the bounds.

We do not find this argument convincing. The empirical error is only a weak predictor of the generalization error, and measuring these errors with different loss functions is a meaningless comparison. The question which hypothesis has lower 0-1-risk cannot be decided on this basis, but only by comparing generalization bounds.

When looking at the bound newly derived above we find that it depends on the sum-loss term $\mathcal{R}_S^{\mathrm{sum}}(f)$. Thus, one may argue that it is a *natural* strategy to minimize this quantity directly instead of the max-loss.

[4] The $\tilde{O}$ (soft O) notation ignores logarithmic factors, not only constant factors. That is, $f(\ell) \in \tilde{O}(g(\ell))$ iff $f(\ell) \in O(g(\ell)\log^{\kappa} g(\ell))$ for some κ.

In general, comparing different machines by means of generalization bounds can be misleading for a number of reasons. The most important is that we are only dealing with upper bounds on the performance, and a *better performance guarantee* does give a *guarantee for better performance*.

5 Discussion

The proposed way to extend generalization bounds for binary large-margin classifiers to large-margin multi-category classifiers is very simple, compared to taking all pairwise interactions between classes into account at once, and it has a number of advantageous properties. It is versatile and generic in the sense that it is applicable to basically every binary margin-based bound. Compared to the underlying bounds we pay the price of considering the worst case over $d(d-1)/2$ pairs of classes. However, also the state-of-the-art results obtained in [9] exhibit the same $\tilde{O}(d^2)$ scaling in the number of classes. In any case this term does not affect the asymptotic tightness of the bounds w.r.t. the number of samples. The same argument, put the other way round, implies that the asymptotic tightness of a bound for binary classification carries over one-to-one to the multi-class case. This implies that binary and multi-class learning have the same sample complexity.

It is straightforward to extend our result to loss functions based on general confusion matrices. Future work may include applying the proposed procedure to more sophisticated bounds for binary classifiers.

Acknowledgements

Christian Igel gratefully acknowledges support from the European Commission through project AKMI (PCIG10-GA-2011-303655).

References

1. Vapnik, V.: Statistical Learning Theory. John Wiley and Sons (1998)
2. Schölkopf, B., Smola, A.J.: Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond. MIT Press (2002)
3. Shawe-Taylor, J., Cristianini, N.: Kernel Methods for Pattern Analysis. Cambridge University Press (2004)
4. Boucheron, S., Bousquet, O., Lugosi, G.: Theory of classification: A survey of some recent advances. ESAIM: Probability and Statistics **9** (2005) 323–375
5. Bartlett, P.L., Jordan, M.I., McAuliffe, J.D.: Convexity, classification, and risk bounds. Journal of the American Statistical Association **101** (2006) 138–156
6. Wu, Q., Ying, Y., Zhou, D.X.: Multi-kernel regularized classifiers. Journal of Complexity **23** (2007) 108–134
7. Steinwart, I., Scovel, C.: Fast rates for support vector machines using Gaussian kernels. The Annals of Statistics **35** (2007) 575–607
8. Steinwart, I.: Oracle inequalities for svms that are based on random entropy numbers. Journal of Complexity **25** (2009) 437–454

9. Guermeur, Y.: VC theory for large margin multi-category classifiers. Journal of Machine Learning Research **8** (2007) 2551–2594
10. Guermeur, Y.: Sample complexity of classifiers taking values in $\mathbb{R}^Q$, Application to multi-class SVMs. Communications in Statistics: Theory and Methods **39** (2010) 543–557
11. Bordes, A., Usunier, N., Bottou, L.: Sequence labelling SVMs trained in one pass. In Daelemans, W., Goethals, B., Morik, K., eds.: Proceedings of the 2008 European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECML-PKDD 2008). Volume 5211 of LNCS., Springer (2008) 146–161
12. Aronszajn, N.: Theory of reproducing kernels. Transactions of the American Mathematical Society **68** (1950) 337–404
13. Boser, B.E., Guyon, I.M., Vapnik, V.N.: A training algorithm for optimal margin classifiers. In: Proceedings of the Fifth Annual Workshop on Computational Learning Theory (COLT 1992), ACM (1992) 144–152
14. Cortes, C., Vapnik, V.: Support-vector networks. Machine Learning **20** (1995) 273–297
15. Weston, J., Watkins, C.: Support vector machines for multi-class pattern recognition. In Verleysen, M., ed.: Proceedings of the Seventh European Symposium On Artificial Neural Networks (ESANN), Evere, Belgium: d-side publications (1999) 219–224
16. Bredensteiner, E.J., Bennett, K.P.: Multicategory classification by support vector machines. Computational Optimization and Applications **12** (1999) 53–79
17. Crammer, K., Singer, Y.: On the algorithmic implementation of multiclass kernel-based vector machines. Journal of Machine Learning Research **2** (2002) 265–292
18. Tsochantaridis, I., Joachims, T., Hofmann, T., Altun, Y.: Large margin methods for structured and interdependent output variables. Journal of Machine Learning Research **6** (2005) 1453–1484