Journal of Computer and System Sciences 82 (2016) 1223–1231

Contents lists available at ScienceDirect

Journal of Computer and System Sciences

www.elsevier.com/locate/jcss

Multi-category classifiers and sample width

Martin Anthony [a], Joel Ratsaby [b,*]

[a] Department of Mathematics, The London School of Economics and Political Science, Houghton Street, London WC2A2AE, U.K.
[b] Electrical and Electronics Engineering Department, Ariel University, Ariel 40700, ISRAEL

ARTICLE INFO

Article history:
Received 31 May 2015
Received in revised form 21 April 2016
Accepted 23 April 2016
Available online 23 June 2016

Keywords:
Multi-category classification
Generalization error
Machine learning
Pattern recognition

ABSTRACT

In a recent paper, the authors introduced the notion of *sample width* for binary classifiers defined on the set of real numbers. It was shown that the performance of such classifiers could be quantified in terms of this sample width. This paper considers how to adapt the idea of sample width so that it can be applied in cases where the classifiers are multi-category and are defined on some arbitrary metric space.

1. Introduction

By a (multi-category) classifier on a set X, we mean a function mapping from X to $[C] = \{1, 2, \ldots, C\}$ where $C \geq 2$ is the number of possible categories. Such a classifier indicates to which of the C different classes objects from X belong and, in supervised machine learning, it is arrived at on the basis of a *sample*, a set of objects from X together with their classifications in $[C]$. In [4], the notion of *sample width* for binary classifiers ($C = 2$) mapping from the real line $X = \mathbb{R}$ was introduced and in [5], this was generalized to finite metric spaces. In this paper, we consider how a similar approach might be taken to the situation in which C could be larger than 2, and in which the classifiers map not simply from the real line, but from some metric space (which would not generally have the linear structure of the real line). The results of this paper are applicable to machine learning, as has been shown recently in [7] for learning case-based inference.

The definition of sample width is given below, but it is possible to indicate the basic idea at this stage: we define sample width to be at least γ if the classifier achieves the correct classifications on the sample and, furthermore, for each sample point, the minimum distance to a point of the domain having a different classification is at least γ.

A key issue that arises in machine learning is that of *generalization error*: given that a classifier has been produced by some learning algorithm on the basis of a (random) sample of a certain size, how can we quantify the accuracy of that classifier, where by its accuracy we mean its likely performance in classifying objects from X correctly? In this paper, we seek answers to this question that involve not just the sample size, but the sample width.

2. Probabilistic modeling of learning

We work in a version of the popular 'PAC' framework of computational learning theory (see [16,9]). This model assumes that the sample **s** consists of an ordered set (x_i, y_i) of labeled examples, where $x_i \in X$ and $y_i \in Y = [C]$, and that each (x_i, y_i)

[*] Corresponding author.
E-mail addresses: m.anthony@lse.ac.uk (M. Anthony), ratsaby@ariel.ac.il (J. Ratsaby).

http://dx.doi.org/10.1016/j.jcss.2016.04.003

in the training sample $\mathbf{s}$ has been generated randomly according to some fixed (but unknown) probability distribution P on $Z = X \times Y$. (This includes, as a special case, the situation in which each x_i is drawn according to a fixed distribution on X and is then labeled deterministically by $y_i = t(x_i)$ where t is some fixed function.) Thus, a sample $\mathbf{s}$ of length m can be thought of as being drawn randomly according to the product probability distribution P^m. An appropriate measure of how well $h : X \to Y$ would perform on further randomly drawn points is its *error*, $\text{er}_P(h)$, the probability that $h(\mathsf{X}) \neq \mathsf{Y}$ for random (X, Y).

Given a function $h \in H$, we can assess how well h fits a training sample through the *sample error*

$$\text{er}_{\mathbf{s}}(h) = \frac{1}{m}|\{i : h(x_i) \neq y_i\}|.$$

This is simply the fraction of sample points not classified correctly by h. Much research in learning theory (see [9,16], for instance) focused on relating the error of a classifier to its sample error, obtaining bounds of the form: for all $\delta \in (0, 1)$, with probability at least $1 - \delta$, for all h belonging to some specified set of functions, $\text{er}_P(h) < \text{er}_{\mathbf{s}}(h) + \epsilon(m, \delta)$, where $\epsilon(m, \delta)$ (known as a *generalization error bound*) is decreasing in m and δ. Such results can be obtained using uniform convergence theorems from probability theory [17,13,10,17,9,16,2]. More recently, emphasis has been placed on 'large-margin' learning (see, for instance [15,2,1,14]) where the idea, in the two-category case, is that if a binary classifier can be thought of as a geometrical separator between points and if it achieves a 'definitive' separation between the points of different classes, then it is a good classifier, and it is possible that a better generalization error bound can be obtained. Margin-based results apply when the binary classifiers are derived from real-valued function by 'thresholding' (taking their sign). Margin analysis has been extended to multi-category classifiers in [11].

3. The width of a classifier

We now discuss the case where the underlying set of objects X forms a metric space. Let X be a set on which is defined a metric $d : X \times X \to \mathbb{R}$. For a subset S of X, define the distance $d(x, S)$ from $x \in X$ to S as follows:

$$d(x, S) := \inf_{y \in S} d(x, y).$$

We define the *diameter* of X to be

$$\text{diam}(X) := \sup_{x, y \in X} d(x, y).$$

We will denote by $\mathcal{H}$ the set of all possible functions h from X to $[C]$.

The paper [4] introduced the notion of the width of a binary classifier at a point in the domain, in the case where the domain was the real line $\mathbb{R}$. Consider a set of points $\{x_1, x_2, \ldots, x_m\}$ from $\mathbb{R}$, which, together with their true classifications $y_i \in \{-1, 1\}$, yield a *training sample*

$$\mathbf{s} = \left((x_j, y_j)\right)_{j=1}^{m} = ((x_1, y_1), (x_2, y_2), \ldots, (x_m, y_m)).$$

We say that $h : \mathbb{R} \to \{-1, 1\}$ achieves sample margin at least γ on $\mathbf{s}$ if $h(x_i) = y_i$ for each i (so that h correctly classifies the sample) and, furthermore, h is constant on each of the intervals $(x_i - \gamma, x_i + \gamma)$. It was then possible to obtain generalization error bounds in terms of the sample width. In this paper we use an analogous notion of width to analyze multi-category classifiers defined on a metric space.

For each k between 1 and C, let us denote by S_k^h the sets corresponding to the function $h : X \to [C]$, defined as follows:

$$S_k^h := h^{-1}(k) = \{x \in X : h(x) = k\}. \tag{1}$$

We define the *width* $w_h(x)$ of h at a point $x \in X$ as follows:

$$w_h(x) := \min_{l \neq h(x)} d(x, S_l^h).$$

In other words, it is the distance from x to the set of points that are labeled differently from $h(x)$. The term 'width' is appropriate since the functional value is just the geometric distance between x and the complement of $S_{h(x)}^h$.

Given $h : X \to [C]$, for each k between 1 and C, we define $f_k^h : X \to \mathbb{R}$ by

$$f_k^h(x) = \min_{l \neq k} d(x, S_l^h) - d(x, S_k^h),$$

and we define $f^h : X \to \mathbb{R}^C$ by setting the kth component function of f^h to be f_k^h: that is, $(f^h)_k = f_k^h$.

Note that if $h(x) = k$, then $f_k^h(x) \geq 0$ and $f_j^h(x) \leq 0$ for $j \neq k$. The function f contains geometrical information encoding how 'definitive' the classification of a point is: if $f_k^h(x)$ is a large positive number, then the point x belongs to category k

and is a large distance from differently classified points. We will regard h as being in error on $(x, y) \in X \times [C]$ if $f_y^h(x)$ is negative. Denoting by X, Y random variables on X and $[C]$, respectively, with a joint probability function P, the error $\mathrm{er}_P(h)$ of h can then be expressed in terms of the function f^h:

$$\mathrm{er}_P(h) = P\left(f_{\mathsf{Y}}^h(\mathsf{X}) < 0\right). \tag{2}$$

We define the class $\mathcal{F}$ of functions as

$$\mathcal{F} := \left\{f^h(x) : h \in \mathcal{H}\right\}. \tag{3}$$

Note that f^h is a mapping from X to the bounded set $[-\mathrm{diam}(X), \mathrm{diam}(X)]^C \subseteq \mathbb{R}^C$. Henceforth, we will use $\gamma > 0$ to denote a *width parameter* whose value is in the range $(0, \mathrm{diam}(X)]$.

For a positive width parameter $\gamma > 0$ and a training sample $\mathbf{s}$, the *empirical* (sample) γ-width error is defined as

$$E_{\mathbf{s}}^{\gamma}(h) := E_{\mathbf{s}}^{\gamma}(f^h) = \frac{1}{m}\sum_{j=1}^{m} \mathbb{I}\left(f_{y_j}^h(x_j) \le \gamma\right). \tag{4}$$

(Here, $\mathbb{I}(A)$ is the indicator function of the set, or event, A.) Note that

$$\begin{aligned} f_y^h(x) \le \gamma &\iff \min_{l \ne y} d(x, S_l^h) - d(x, S_y^h) \le \gamma \\ &\iff \exists l \ne y \text{ such that } d(x, S_l^h) \le d(x, S_y^h) + \gamma. \end{aligned}$$

So the empirical γ-width error on the sample is the proportion of points in the sample which are either misclassified by h or which are classified correctly, but lie within distance γ of the set of points classified differently. (We recall that $h(x) = y$ implies $d(x, S_y^h) = 0$.) Our aim is to show that (with high probability) the generalization error $\mathrm{er}_P(h)$ is not much greater than $E_{\mathbf{s}}^{\gamma}(h)$. (In particular, as a special case, we want to bound the generalization error given that $E_{\mathbf{s}}^{\gamma}(h) = 0$.) This will imply that if the learner finds a hypothesis which, for a large value of γ, has a small γ-width error, then that hypothesis is likely to have small error. What this indicates, then, is that if a hypothesis has a large width on most points of a sample, then it will be likely to have small error.

4. Covering numbers

4.1. Covering numbers

A central idea in large-margin analysis is that of *covering number* and this will also prove useful here. We will discuss different types of covering numbers, so we introduce the idea in some generality to start with.

Suppose (A, d) is a (pseudo-)metric space and that $\alpha > 0$. Then an α-cover of A (with respect to d) is a finite set C (possibly a subset of A) such that, for every $a \in A$, there is some $c \in C$ such that $d(a, c) \le \alpha$. If such a cover exists, then the minimum cardinality of such a cover is the *covering number* $\mathcal{N}(A, \alpha, d)$.

We are working with the set $\mathcal{F}$ of vector-valued functions from X to $\mathbb{R}^C$, as defined earlier. We define the sup-metric d_∞ on F as follows: for $f, g : X \to \mathbb{R}^C$,

$$d_\infty(f, g) = \sup_{x \in X} \max_{1 \le k \le C} |f_k(x) - g_k(x)|,$$

where f_k denotes the kth component function of f. (Note that each component function is bounded, so the metric is well-defined.)

We can bound the covering numbers $\mathcal{N}(\mathcal{F}, \alpha, d_\infty)$ of $\mathcal{F}$ (with respect to the sup-metric) in terms of the covering numbers of X with respect to its metric d. The result is as follows.

Theorem 4.1. *For $\alpha \in (0, \mathrm{diam}(X)]$,*

$$\mathcal{N}(\mathcal{F}, \alpha, d_\infty) \le \left(\frac{9\,\mathrm{diam}(X)}{\alpha}\right)^{CN_\alpha},$$

where $N_\alpha = \mathcal{N}(X, \alpha/3, d)$.

4.2. Smoothness of the function class

As a first step towards establishing this result, we prove that the functions in $\mathcal{F}$ satisfy a certain Lipschitz (or smoothness) property. A similar property was proved for the case of a binary classifier on a finite metric space in [5]: this generalizes that result to the multi-category case, and deals with the case in which the underlying metric space may be infinite.

Proposition 4.2. *For every $f \in \mathcal{F}$, and for all $x, x' \in X$,*

$$\max_{1 \le k \le C} |f_k(x) - f_k(x')| \le 2\, d(x, x'). \tag{5}$$

Proof. Let $x, x' \in X$ and fix k between 1 and C. We show that

$$|f_k(x) - f_k(x')| \le 2\, d(x, x').$$

Recall that, since $f \in \mathcal{F}$, there is some $h : X \to [C]$ such that, for all x,

$$f_k(x) = \min_{l \ne k} d(x, S_l^h) - d(x, S_k^h)$$

where, for each i, $S_i^h = h^{-1}(i)$. We have

$$\begin{aligned} |f_k(x) - f_k(x')| &= \left| \min_{l \ne k} d(x, S_l^h) - d(x, S_k^h) - \min_{l \ne k} d(x', S_l^h) + d(x', S_k^h) \right| \\ &\le \left| \min_{l \ne k} d(x, S_l^h) - \min_{l \ne k} d(x', S_l^h) \right| + \left| d(x, S_k^h) - d(x', S_k^h) \right| \end{aligned}$$

We consider in turn each of the two terms in this final expression. We start with the second, by showing that, for any set S, $|d(x, S) - d(x', S)| \le d(x, x')$. From the fact that, for each $s \in S$, $d(x, s) \le d(x, x') + d(x', s)$, it follows that

$$\inf_{s \in S} d(x, s) \le d(x, x') + \inf_{s \in S} d(x', s);$$

that is,

$$d(x, S) \le d(x, x') + d(x', S).$$

An analogous argument with x, x' interchanged establishes

$$d(x', S) \le d(x, x') + d(x, S).$$

Next we show

$$\left| \min_{l \ne k} d(x, S_l^h) - \min_{l \ne k} d(x', S_l^h) \right| \le d(x, x').$$

Suppose that $\min_{k \ne l} d(x, S_l^h) = d(x, S_p^h)$ and that $\min_{k \ne l} d(x', S_l^h) = d(x', S_q^h)$. Then,

$$d(x, S_p^h) \le d(x, S_q^h) \le d(x, x') + d(x', S_q^h)$$

and

$$d(x', S_q^h) \le d(x', S_p^h) \le d(x, x') + d(x, S_p^h).$$

It follows that $|d(x, S_p^h) - d(x', S_q^h)| \le d(x, x')$. □

Next, we exploit this ‘smoothness’ to construct a cover for $\mathcal{F}$.

4.3. Covering $\mathcal{F}$

Let the subset $C_\alpha \subseteq X$ be a *minimal* size $\alpha/3$-cover for X with respect to the metric d. So, for every $x \in X$ there is some $\hat{x} \in C_\alpha$ such that $d(x, \hat{x}) \leq \alpha/3$. Denote by N_α the cardinality of C_α.

Let

$$\Lambda_\alpha = \left\{\lambda_i = i\alpha : i = -\left\lceil \frac{3\,\mathrm{diam}(X)}{\alpha} \right\rceil, \ldots, -1, 0, 1, 2, \ldots, \left\lceil \frac{3\,\mathrm{diam}(X)}{\alpha} \right\rceil\right\} \tag{6}$$

and define the class $\hat{F}$ to be all functions $\hat{f} : C_\alpha \to (\Lambda_\alpha)^C$. Then $\hat{F}$ is of a finite size equal to $|\Lambda_\alpha|^{C\,N_\alpha}$. For any $\hat{f} \in \hat{F}$ define the extension $\hat{f}_{ext} : X \to \mathbb{R}^C$ of $\hat{f}$ to the whole domain X as follows: given $\hat{f}$ (which is well-defined on the points $\hat{x}_i$ of the cover) then for every point x in the ball $B_{\alpha/3}(\hat{x}_i) = \{x \in X : d(x, \hat{x}_i) \leq \alpha/3\}$, we let $\hat{f}_{ext}(x) = \hat{f}(\hat{x}_i)$, for all $\hat{x}_i \in C_\alpha$ (where, if, for a point x there is more than one point $\hat{x}_i$ such that $x \in B_{\alpha/3}(\hat{x}_i)$, we arbitrarily pick one of the points $\hat{x}_i$ in order to assign the value of $\hat{f}_{ext}(x)$). There is a one-to-one correspondence between the functions $\hat{f}$ and the functions $\hat{f}_{ext}$. Hence the set $\hat{F}_{ext} = \left\{\hat{f}_{ext} : \hat{f} \in \hat{F}\right\}$ is of cardinality equal to $|\Lambda_\alpha|^{C\,N_\alpha}$.

We claim that for any $f \in \mathcal{F}$ there exists an $\hat{f}_{ext}$ such that $d_\infty(f, \hat{f}_{ext}) \leq \alpha$. To see this, first for every point $\hat{x}_i \in C_\alpha$, consider $f(\hat{x}_i)$ and find a corresponding element in Λ_α^C, (call it $\hat{f}(\hat{x}_i)$) such that

$$\max_{1 \leq k \leq C} |(f(\hat{x}_i))_k - (\hat{f}(\hat{x}_i))_k| \leq \alpha/3. \tag{7}$$

(That there exists such a value follows by design of Λ_α.) By the above definition of extension, it follows that for all points $x \in B_{\alpha/3}(\hat{x}_i)$ we have $\hat{f}_{ext}(x) = \hat{f}(\hat{x}_i)$. Now, from (5) we have for all $f \in \mathcal{F}$,

$$\max_{1 \leq i \leq k} \sup_{x \in B_{\alpha/3}(\hat{x}_i)} |(f(x))_k - (f(\hat{x}_i))_k| \leq 2d(x, \hat{x}_i) \leq 2\alpha/3. \tag{8}$$

Hence for any $f \in \mathcal{F}$ there exists a function $\hat{f} \in \hat{F}$ with a corresponding $\hat{f}_{ext} \in \hat{F}_{ext}$ such that, given an $x \in X$, there exists $\hat{x}_i \in C_\alpha$ such that, for each k between 1 and C, $|(f(x))_k - (\hat{f}_{ext}(x))_k| = |(f(x))_k - (\hat{f}_{ext}(\hat{x}_i))_k|$. The right hand side can be expressed as

$$\begin{aligned}
|(f(x))_k - (\hat{f}_{ext}(\hat{x}_i))_k| &= |(f(x))_k - (\hat{f}(\hat{x}_i))_k| \\
&= |(f(x))_k - (f(\hat{x}_i))_k + (f(\hat{x}_i))_k - (\hat{f}(\hat{x}_i))_k| \\
&\leq |(f(x))_k - (f(\hat{x}_i))_k| + |(f(\hat{x}_i))_k - (\hat{f}(\hat{x}_i))_k| \\
&\leq 2\alpha/3 + \alpha/3 \qquad (9) \\
&= \alpha,
\end{aligned}$$

where (9) follows from (7) and (8).

Hence the set $\hat{F}_{ext}$ forms an α-covering of the class $\mathcal{F}$ in the sup-norm. Thus we have the following covering number bound.

$$\mathcal{N}(\mathcal{F}, \alpha, d_\infty) \leq |\Lambda_\alpha|^{C\,N_\alpha} = \left(2\left\lceil \frac{3\,\mathrm{diam}(X)}{\alpha} \right\rceil + 1\right)^{C\,N_\alpha}. \tag{10}$$

Theorem 4.1 now follows because (for $0 < \alpha \leq \mathrm{diam}(X)$)

$$2\left\lceil \frac{3\,\mathrm{diam}(X)}{\alpha} \right\rceil + 1 \leq 2\left(\frac{3\,\mathrm{diam}(X)}{\alpha} + 1\right) + 1 = \frac{6\,\mathrm{diam}(X)}{\alpha} + 3 \leq \frac{9\,\mathrm{diam}(X)}{\alpha}.$$

5. Generalization error bounds

We present two results. The first bounds the generalization error in terms of a width parameter γ for which the γ-width error on the sample is zero; the second (more general but looser in that special case) bounds the error in terms of γ and the γ-width error on the sample (which could be non-zero).

Theorem 5.1. *Suppose that X is a metric space of diameter* $\mathrm{diam}(X)$. *Suppose P is any probability measure on $Z = X \times [C]$. Let $\delta \in (0, 1)$. Then, with probability at least $1 - \delta$, the following holds for $\mathbf{s} \in Z^m$: for any function $h : X \to [C]$, and for any $\gamma \in (0, \mathrm{diam}(X)]$, if $E_{\mathbf{s}}^{\gamma}(h) = 0$, then*

$$\mathrm{er}_P(h) \leq \frac{2}{m}\left(C\mathcal{N}(X, \gamma/12, d)\log_2\left(\frac{36\,\mathrm{diam}(X)}{\gamma}\right) + \log_2\left(\frac{4\,\mathrm{diam}(X)}{\delta\gamma}\right)\right).$$

Theorem 5.2. *Suppose that X is a metric space of diameter* $\mathrm{diam}(X)$. *Suppose P is any probability measure on $Z = X \times [C]$. Let $\delta \in (0,1)$. Then, with probability at least $1-\delta$, the following holds for $\mathbf{s} \in Z^m$: for any function $h : X \to [C]$, and for any $\gamma \in (0, \mathrm{diam}(X)]$,*

$$\mathrm{er}_P(h) \le E_{\mathbf{s}}^{\gamma}(h) + \sqrt{\frac{2}{m}\left(C\mathcal{N}(X,\gamma/12,d)\ln\left(\frac{36\,\mathrm{diam}(X)}{\gamma}\right) + \ln\left(\frac{4\,\mathrm{diam}(X)}{\gamma\delta}\right)\right)} + \frac{1}{m}.$$

What we have in Theorem 5.2 is a high probability bound that takes the following form: for all h and for all $\gamma \in (0, \mathrm{diam}(X)]$,

$$\mathrm{er}_P(h_S) \le E_{\mathbf{s}}^{\gamma}(h) + \epsilon(m,\gamma,\delta),$$

where ϵ tends to 0 as $m \to \infty$ and ϵ decreases as γ increases. The rationale for seeking such a bound is that there is likely to be a trade-off between width error on the sample and the value of ϵ: taking γ small so that the error term $E_{\mathbf{s}}^{\gamma}(h)$ is zero might entail a large value of ϵ; and, conversely, choosing γ large will make ϵ relatively small, but lead to a large sample error term. So, in principle, since the value γ is free to be chosen, one could optimize the choice of γ on the right-hand side of the bound to minimize it.

The bound of Theorem 5.2 compares well with the margin-based bound from [11]. It varies as $1/\sqrt{m}$, while that of [11] has an additional $\sqrt{\ln m}$ factor. The dependence on C is, however, $\sqrt{C}$ whereas in [11] it is $\sqrt{\ln^2 C}$. The bounds are not directly comparable because ours concerns width and that of [11] involves margin, but, for fixed C, it is notable that the m-dependence of our bound is better. The dependence of Theorem 5.2 on the width parameter γ is, in general, similar to the dependence of [11] on the margin parameter γ as both grow like $\sqrt{\mathcal{N}_\gamma}$ where $\mathcal{N}_\gamma$ is the covering number of the underlying real-valued discriminant function class. The advantage of the bound of Theorem 5.2 is that it is expressed in terms of the covering number of the actual metric space which, in some problems, such as when the metric space is finite [5], can be efficiently estimated.

Proof of Theorem 5.1. The proof uses techniques similar to those first used in [17,16,10,13] and in subsequent work extending those techniques to learning with real-valued functions, such as [12,3,1,8,6]. The first observation is that if

$$Q = \{\mathbf{s} \in Z^m : \exists h \in \mathcal{H} \text{ with } E_{\mathbf{s}}^{\gamma}(h) = 0,\ \mathrm{er}_P(h) \ge \epsilon\}$$

and

$$T = \{(\mathbf{s},\mathbf{s}') \in Z^m \times Z^m : \exists h \in \mathcal{H} \text{ with } E_{\mathbf{s}}^{\gamma}(h) = 0,\ E_{\mathbf{s}'}^{0}(h) \ge \epsilon/2\},$$

then, for $m \ge 8/\epsilon$,

$$P^m(Q) \le 2\,P^{2m}(T).$$

This follows from the proof of Lemma 10.2 in [1]; instead of the γ-margin error event defined there as $\mathrm{Y}f(\mathrm{X}) < \gamma$ based on *any* real-valued function f on X (the empirical γ-error of f is denoted by $\hat{\mathrm{er}}_{\mathbf{s}}^{\gamma}(f)$), the current paper considers the γ-width error event which is defined as $f_{\mathrm{Y}}^{h}(\mathrm{X}) < \gamma$ and is based on the specific real valued function, the width function, f^h. In the proof of that lemma, substitute for f the width function f^h, set the value of $\hat{\mathrm{er}}_r^{\gamma}(f^h) = 0$ and apply Chebyshev's inequality to show that if $\mathrm{er}_P(f^h) \ge \epsilon$, then for $m \ge 8/\epsilon$, $P^m(\hat{\mathrm{er}}_{\mathbf{s}}(f^h) \ge \epsilon/2) \ge 1/2$, for any h.

Let G be the permutation group (the 'swapping group') on the set $\{1,2,\ldots,2m\}$ generated by the transpositions $(i, m+i)$ for $i = 1,2,\ldots,m$. Then G acts on Z^{2m} by permuting the coordinates: for $\sigma \in G$,

$$\sigma(z_1,z_2,\ldots,z_{2m}) = (z_{\sigma(1)},\ldots,z_{\sigma(2m)}).$$

From the proof of Lemma 10.2 in [1], by invariance of P^{2m} under the action of G, we have

$$P^{2m}(T) \le \max\{\mathbb{P}(\sigma\mathbf{z} \in T) : \mathbf{z} \in Z^{2m}\},$$

where $\mathbb{P}$ denotes the probability over uniform choice of σ from G.

Let $\mathcal{F} = \{f^h : h \in \mathcal{H}\}$ be the set of vector-valued functions derived from $\mathcal{H}$ as before, and let $\hat{\mathcal{F}}$ be a minimal $\gamma/2$-cover of $\mathcal{F}$ in the d_∞-metric. Theorem 4.1 tells us that the size of $\hat{\mathcal{F}}$ is no more than

$$\left(\frac{18\,\mathrm{diam}(X)}{\gamma}\right)^{CN},$$

where $N = \mathcal{N}(X,\gamma/6,d)$.

The next part of the argument is similar to that of Theorem 2.2 in [6], which follows earlier 'symmetrization' proofs. We omit most of the details. It can be shown that, for any $\mathbf{z}$,

$$\mathbb{P}(\sigma\mathbf{z}\in T)\le\sum_{\hat f\in\hat{\mathcal{F}}}\mathbb{P}(\sigma\mathbf{z}\in T(\hat f))\le|\hat{\mathcal{F}}|2^{-\epsilon m/2},$$

where, for $\hat f\in\hat F$,

$$T(\hat f):=\{(\mathbf{s},\mathbf{s}')\in Z^m\times Z^m: E_{\mathbf{s}}^{\gamma/2}(\hat f)=0,\ E_{\mathbf{s}'}^{\gamma/2}(\hat f)\ge\epsilon/2\}.$$

So,

$$P^m(Q)\le 2\,P^{2m}(T)\le 2\,|\hat{\mathcal{F}}|\,2^{-\epsilon m/2}\le 2\left(\frac{18\,\mathrm{diam}(X)}{\gamma}\right)^{CN},$$

where $N=\mathcal{N}(X,\gamma/6,d)$. This is at most δ when

$$\epsilon=\frac{2}{m}\left(CN\log_2\left(\frac{18\,\mathrm{diam}(X)}{\gamma}\right)+\log_2\left(\frac{2}{\delta}\right)\right).$$

Next, we use this to obtain a result in which γ is not prescribed in advance. For $\alpha_1,\alpha_2,\delta\in(0,1)$, let $E(\alpha_1,\alpha_2,\delta)$ be the set of $\mathbf{z}\in Z^m$ for which there exists some $h\in\mathcal{H}$ with $E_{\mathbf{z}}^{\alpha_2}(h)=0$ and $\mathrm{er}_P(h)\ge\epsilon_1(m,\alpha_1,\delta)$, where

$$\epsilon_1(m,\alpha_1,\delta)=\frac{2}{m}\left(C\mathcal{N}(X,\alpha_1/6,d)\log_2\left(\frac{18\,\mathrm{diam}(X)}{\alpha_1}\right)+\log_2\left(\frac{2}{\delta}\right)\right).$$

Then the result just obtained tells us that $P^m(E(\alpha,\alpha,\delta))\le\delta$. It is also clear that if $\alpha_1\le\alpha\le\alpha_2$ and $\delta_1\le\delta$, then $E(\alpha_1,\alpha_2,\delta_1)\subseteq E(\alpha,\alpha,\delta)$. Let D denote $\mathrm{diam}(X)$. Then, following an argument from [8],

$$E\left(\gamma/2,\gamma,\delta\gamma/2D\right)\subseteq E\left(\frac{D}{2^{l+1}},\frac{D}{2^{l+1}},\frac{\delta}{2^{l+1}}\right),$$

for all γ satisfying

$$\frac{D}{2^{l+1}}\le\gamma\le\frac{D}{2^l},$$

and therefore

$$\begin{aligned}
&P^m\left(\bigcup_{\gamma\in(0,D]}E\left(\gamma/2,\gamma,\delta\gamma/2D\right)\right)\\
&=P^m\left(\bigcup_{l=0}^{\infty}\ \bigcup_{D/2^{l+1}\le\gamma\le D/2^l}E\left(\frac{D}{2^{l+1}},\frac{D}{2^{l+1}},\frac{\delta}{2^{l+1}}\right)\right)\\
&\le\sum_{l=0}^{\infty}P^m\left(E\left(\frac{D}{2^{l+1}},\frac{D}{2^{l+1}},\frac{\delta}{2^{l+1}}\right)\right)\\
&\le\delta\sum_{l=0}^{\infty}(1/2^l)\\
&\le\delta
\end{aligned}$$

In other words, with probability at least $1-\delta$, *for all* $\gamma\in(0,\mathrm{diam}(X)]]$, we have that if $h\in\mathcal{H}$ satisfies $E_{\mathbf{s}}^{\gamma}(h)=0$, then $\mathrm{er}_P(h)<\epsilon_2(m,\gamma,\delta)$, where

$$\epsilon_2(m,\gamma,\delta)=\frac{2}{m}\left(C\mathcal{N}(X,\gamma/12,d)\log_2\left(\frac{36\,\mathrm{diam}(X)}{\gamma}\right)+\log_2\left(\frac{4\,\mathrm{diam}(X)}{\delta\gamma}\right)\right).$$

Note that γ now need not be prescribed in advance. □

Proof of Theorem 5.2. Guermeur [11] has developed a framework in which to analyze multi-category classification, and we can apply one of his results to obtain the bound of Theorem 5.2, a generalization error bound applicable to the case in which the γ-width sample error is not zero. In that framework, there is a set $\mathcal{G}$ of functions from X into $\mathbb{R}^C$, and a typical $g\in\mathcal{G}$ is represented by its component functions g_k for $k=1$ to C. Each $g\in\mathcal{G}$ satisfies the constraint

$$\sum_{k=1}^{C}g_k(x)=0,\ \ \forall x\in X.$$

The *risk* $R(g)$ of $g \in \mathcal{G}$, when the underlying probability measure on $X \times Y$ is P, is defined to be the P-probability that for $(\mathsf{X}, \mathsf{Y}) \in X \times [C]$, we have $g_{\mathsf{Y}}(\mathsf{X}) \le \max_{k \neq \mathsf{Y}} g_k(\mathsf{X})$. For $g \in \mathcal{G}$, Δg is defined to be the function $X \to \mathbb{R}^C$ given by

$$(\Delta g)_k(x) = \frac{1}{2}\left(g_k(x) - \max_{l \neq k} g_l(x)\right), \quad 1 \le k \le C.$$

We define the class of such functions by

$$\Delta\mathcal{G} := \{\Delta g : g \in \mathcal{G}\}. \tag{11}$$

Given a sample $\mathbf{s} \in (X \times [C])^m$, let

$$R_{\gamma,\mathbf{s}}(g) = \frac{1}{m}\sum_{i=1}^{m} \mathbb{I}\left\{\Delta g_{y_i}(x_i) < \gamma\right\}.$$

Then a result following from [11] is (in the above notation) as follows:

Let $\delta \in (0,1)$ and suppose P is a probability measure on $Z = X \times [C]$. With P^m-probability at least $1-\delta$, $\mathbf{s} \in Z^m$ will be such that we have the following: (for any fixed $d > 0$) for all $\gamma \in (0,d]$ and for all $g \in \mathcal{G}$,

$$R(g) \le R_{\gamma,\mathbf{s}}(g) + \sqrt{\frac{2}{m}\left(\ln \mathcal{N}(\Delta\mathcal{G}, \gamma/4, d_\infty) + \ln\left(\frac{2d}{\gamma\delta}\right)\right)} + \frac{1}{m}.$$

(In fact, the result from [11] involves empirical covering numbers rather than d_∞-covering numbers. The latter are at least as large as the empirical covering numbers, but we use these because we have bounded them earlier in this paper.)

We can (as in [6]) formulate our problem in Guermeur's framework and involve the functions f^h from earlier. For each function $h : X \to [C]$, let

$$g^h : X \to \mathbb{R}^C$$

be given by

$$g_k^h(x) = \frac{1}{C}\sum_{i=1}^{C} d(x, S_i^h) - d(x, S_k^h),$$

where, as before, $S_j^h = h^{-1}(j)$. Let

$$\mathcal{G} = \{g^h : h \in \mathcal{H}\}$$

be the set of all such g and take $\Delta\mathcal{G}$ as in (11). Then these functions satisfy the constraint that their coordinate functions sum to the zero function, since

$$\sum_{k=1}^{C} g_k^h(x) = \sum_{k=1}^{C} \frac{1}{C}\sum_{i=1}^{C} d(x, S_i^h) - \sum_{k=1}^{C} d(x, S_k^h) = \sum_{k=1}^{C} d(x, S_k^h) - \sum_{k=1}^{C} d(x, S_k^h) = 0.$$

Furthermore, for each k,

$$\begin{aligned}\Delta g_k^h(x) &= \frac{1}{2}\left(g_k^h(x) - \max_{l \neq k} g_l^h(x)\right)\\ &= \frac{1}{2}\left(\frac{1}{C}\sum_{i=1}^{C} d(x, S_i^h) - d(x, S_k^h) - \max_{l \neq k}\left(\frac{1}{C}\sum_{i=1}^{C} d(x, S_i^h) - d(x, S_l^h)\right)\right),\end{aligned}$$

which is easily seen to be

$$\frac{1}{2}\left(\min_{l \neq k} d(x, S_l^h) - d(x, S_k^h)\right) = \frac{1}{2} f_k^h(x).$$

From the definition of g^h, the event that $g_{\mathsf{Y}}^h(\mathsf{X}) \le \max_{k \neq \mathsf{Y}} g_k^h(\mathsf{X})$ (the probability of which is, by definition, $R(g^h)$) is equivalent to the event that

$$\frac{1}{C}\sum_{i=1}^{C} d(\mathsf{X}, S_i^h) - d(\mathsf{X}, S_{\mathsf{Y}}^h) \le \max_{k \neq \mathsf{Y}}\left(\frac{1}{C}\sum_{i=1}^{C} d(\mathsf{X}, S_i^h) - d(\mathsf{X}, S_k^h)\right),$$

which is equivalent to $\min_{k \neq \mathsf{Y}} d(\mathsf{X}, S_k^h) \le d(\mathsf{X}, S_{\mathsf{Y}}^h)$. It can therefore be seen that $R(g^h) = \mathrm{er}_P(h)$. Similarly, $R_{\gamma,\mathbf{s}}(g^h) = E_{\mathbf{s}}^{2\gamma}(h)$.

Noting that $\Delta\mathcal{G} = (1/2)\mathcal{F}$, so that an $\alpha/2$ cover of $\mathcal{F}$ will provide an $\alpha/4$ cover of $\Delta\mathcal{G}$, we can therefore apply Guermeur's result to see that with probability at least $1-\delta$, for all h and for all $\gamma \in (0, \mathrm{diam}(X)]$,

$$\begin{aligned}
\mathrm{er}_P(h) &\le E_{\mathbf{s}}^{2\gamma}(h) + \sqrt{\frac{2}{m}\left(\ln \mathcal{N}(\mathcal{F}, \gamma/2, d_\infty) + \ln\left(\frac{2\,\mathrm{diam}(X)}{\gamma\delta}\right)\right)} + \frac{1}{m} \\
&= E_{\mathbf{s}}^{\gamma}(h) + \sqrt{\frac{2}{m}\left(\ln \mathcal{N}(\mathcal{F}, \gamma/4, d_\infty) + \ln\left(\frac{4\,\mathrm{diam}(X)}{\gamma\delta}\right)\right)} + \frac{1}{m} \\
&\le E_{\mathbf{s}}^{\gamma}(h) + \sqrt{\frac{2}{m}\left(C\mathcal{N}(X, \gamma/12, d)\ln\left(\frac{36\,\mathrm{diam}(X)}{\gamma}\right) + \ln\left(\frac{4\,\mathrm{diam}(X)}{\gamma\delta}\right)\right)} + \frac{1}{m}. \qquad \square
\end{aligned}$$

6. Conclusions

This paper generalizes considerably the initial notion of sample width introduced in [4], where the focus was on binary-valued functions defined on the real line. It also extends results from [5], in which binary classification on a finite metric space was studied. (The focus there was on a finite domain so that the covering numbers, and hence generalization error bounds, could be bounded by certain graph-theoretical parameters associated with the underlying metric space.) This paper provides generalization error bounds for any multi-category classifiers on a metric space, and the bounds involve both the covering numbers of the underlying metric space and the extent to which a classifier achieves a large sample width on the training sample. The results of this paper are directly applicable to machine learning, as for instance, to learning case base inference [7].

Acknowledgments

This work was supported in part by the IST Programme of the European Community, under the PASCAL2 Network of Excellence, IST-2007-216886 and by a research grant from the Suntory and Toyota International Centres for Economics and Related Disciplines at the London School of Economics. We are grateful to the referees for their careful reading of the paper and their many helpful comments.

References

[1] M. Anthony, P.L. Bartlett, Function learning from interpolation, Comb. Probab. Comput. 9 (1) (2000) 213–225.
[2] M. Anthony, P.L. Bartlett, Neural Network Learning: Theoretical Foundations, Cambridge University Press, 1999.
[3] N. Alon, S. Ben-David, N. Cesa-Bianchi, D. Haussler, Scale-sensitive dimensions, uniform convergence, and learnability, J. ACM 44 (4) (1997) 615–631.
[4] M. Anthony, J. Ratsaby, Maximal width learning of binary functions, Theor. Comput. Sci. 411 (1) (2010) 138–147.
[5] M. Anthony, J. Ratsaby, Learning bounds via sample width for classifiers on finite metric spaces, Theor. Comput. Sci. 529 (2014) 2–10.
[6] M. Anthony, J. Ratsaby, Analysis of a multi-category classifier, Discrete Appl. Math. 160 (16–17) (2012) 2329–2338.
[7] M. Anthony, J. Ratsaby, A probabilistic approach to case-based inference, Theor. Comput. Sci. 589 (2015) 61–75.
[8] P.L. Bartlett, The sample complexity of pattern classification with neural networks: the size of the weights is more important than the size of the network, IEEE Trans. Inf. Theory 44 (2) (1998) 525–536.
[9] A. Blumer, A. Ehrenfeucht, D. Haussler, M.K. Warmuth, Learnability and the Vapnik–Chervonenkis dimension, J. ACM 36 (4) (1989) 929–965.
[10] R.M. Dudley, Uniform Central Limit Theorems, Camb. Stud. Adv. Math., vol. 63, Cambridge University Press, Cambridge, UK, 1999.
[11] Y. Guermeur, VC theory of large margin multi-category classifiers, J. Mach. Learn. Res. 8 (2007) 2551–2594.
[12] D. Haussler, Decision theoretic generalizations of the pac model for neural net and other learning applications, Inf. Comput. 100 (1) (1992) 78–150.
[13] D. Pollard, Convergence of Stochastic Processes, Springer-Verlag, 1984.
[14] J. Shawe-Taylor, P.L. Bartlett, R.C. Williamson, M. Anthony, Structural risk minimization over data-dependent hierarchies, IEEE Trans. Inf. Theory 44 (5) (1996) 1926–1940.
[15] A.J. Smola, P.L. Bartlett, B. Scholkopf, D. Schuurmans, Advances in Large-Margin Classifiers (Neural Information Processing), MIT Press, 2000.
[16] V.N. Vapnik, Statistical Learning Theory, Wiley, 1998.
[17] V.N. Vapnik, A.Y. Chervonenkis, On the uniform convergence of relative frequencies of events to their probabilities, Theory Probab. Appl. 16 (2) (1971) 264–280.