ARTICLE IN PRESS

Discrete Applied Mathematics ▮ (▮▮▮▮) ▮▮▮–▮▮▮

Contents lists available at ScienceDirect

Discrete Applied Mathematics

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

A hybrid classifier based on boxes and nearest neighbors

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

[a] Department of Mathematics, The London School of Economics and Political Science, Houghton Street, London WC2A2AE, UK
[b] Electrical and Electronics Engineering Department, Ariel University of Samaria, Ariel 40700, Israel

ARTICLE INFO

Article history:
Received 28 November 2011
Received in revised form 30 July 2013
Accepted 16 August 2013
Available online xxxx

Keywords:
Logical analysis of data
LAD methods
Generalization error
Machine learning
Learning algorithms
Large margin learning

ABSTRACT

In this paper we analyse the generalization performance of a type of binary classifier defined on the unit cube. This classifier combines some of the aspects of the standard methods that have been used in the logical analysis of data (LAD) and geometric classifiers, with a nearest-neighbor paradigm. We assess the predictive performance of the new classifier in learning from a sample, obtaining generalization error bounds that improve as a measure of 'robustness' of the classifier on the training sample increases.

1. Introduction

In this paper we study a method of classifying points of $[0, 1]^n$ into two classes. The classifiers we use combine the use of 'boxes' with a nearest-neighbor approach and for this reason we describe it as a *hybrid* classifier. Both classification by boxes and nearest-neighbor classification have been widely used. For instance, the use of boxes is integral to many of the standard methods used in the logical analysis of data (LAD); see [8,9], for instance.

The primary purpose of this paper is to quantify the performance of the hybrid classifiers by bounding their generalization error. In doing so, we obtain bounds that depend on a measure of how 'robust' the classification is on the training sample. In using real-valued functions to form the basis of the classification, we can also attach some degree of 'confidence' or 'definitiveness' to the resulting classifications, and this could be of some practical use.

In Section 2, we give some background by way of motivation. Chiefly, this is a description of some of the standard methods used in the logical analysis of data, especially as they apply to data in which the data points are not necessarily binary, but have real-valued components. In this context, unions of boxes are used as a key means of classification. (It should be noted that classification by unions of boxes has been more widely studied, not just in the context of LAD; see [10] for instance.)

Section 3 describes a type of 'hybrid' classifier which incorporates some of the features of the LAD (and other) techniques in that it uses unions of boxes. However, the classifiers combine this with a nearest-neighbor paradigm for classifying some regions of the domain. In this section, we define the classifiers, give an example, and discuss the rationale for this method of classification.

Section 4 provides the main theoretical results, giving bounds on the predictive performance (or generalization error) of the classifiers. The bounds we obtain are better if the classifier achieves 'definitively' correct classification of the sample points.

* 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.dam.2014.02.018
0166-218X/

Please cite this article in press as: M. Anthony, J. Ratsaby, A hybrid classifier based on boxes and nearest neighbors, Discrete Applied Mathematics (2014), http://dx.doi.org/10.1016/j.dam.2014.02.018

2. Classification using unions of boxes

In standard logical analysis of data (LAD) for binary data, we have some collection of labeled *observations* (or data-points, or training examples) (x_i, b_i), for $i = 1, 2, \ldots, m$, where m is known as the sample size. The observations are the x_i and their labels are the b_i. The $x_i \in \{0, 1\}^n$ for which $(x_i, 1)$ appears among the observations are said to be the *positive observations*; and those for which $(x_i, 0)$ appears are the *negative observations*. We denote the sets of positive and negative observations by D^+ and D^- respectively, and the set of all m observations by D. The primary aim is to find some function $h : \{0, 1\}^n \to \{0, 1\}$, a *hypothesis* or *classifier*, that describes the classifications of the known observations well and therefore, as a result, would act as a reliable guide to how future, as yet unseen, elements of $\{0, 1\}^n$ ought to be classified. This is, indeed, a central issue generally in machine learning. The approach taken in LAD methods involves the use of Boolean functions constructed in precise algorithmic ways from the observations. In the standard LAD method for binary data [11], a disjunctive normal form Boolean function (a DNF) is produced. The terms of this DNF are called *positive patterns*. A (pure) positive pattern is a conjunction of literals which is true on at least one positive observation (in which case we say that the observation is *covered* by the pattern) but which is not true on any negative observation. The classifier is then taken to be the disjunction of a set of positive patterns. A more general technique combines the use of positive patterns with *negative* patterns, conjunctions which cover some negative observations. Points of $\{0, 1\}^n$ are then classified as follows: $x \in \{0, 1\}^n$ is assigned value 1 if it is covered by at least one positive pattern, but no negative patterns; and it is assigned value 0 if it is covered by at least one negative pattern, but no positive patterns. If a point x is covered by both types of pattern (which might well be the case, even if we have been careful to ensure that the observations themselves have only been covered by patterns of one type) then its classification is often determined by using a *discriminant*, which takes into account (perhaps in a weighted way) the number of positive and the number of negative patterns covering it.

These standard LAD techniques apply when the data is binary. However, many applications involve numerical data, in which $D \subseteq [0, 1]^n \times \{0, 1\}$. The LAD methods have been extended to deal with such cases; see [9], for instance. The approach is first to *binarize* the data, so that observations $x \in [0, 1]^n$ are converted into binary observations $x^* \subseteq \{0, 1\}^d$, where, generally, $d \geq n$. The standard way to do so is to use *cutpoints* for each attribute (that is, for each of the n geometrical dimensions). For each coordinate (or dimension) $j = 1, 2, \ldots, n$, let $u_1^{(j)}, u_2^{(j)}, \ldots, u_{k_j}^{(j)}$ be, in increasing order, all the distinct values of the jth coordinate of the observations in D. For each j, let

$$\beta_i^{(j)} = \frac{u_i^{(j)} + u_{i+1}^{(j)}}{2}$$

for $i = 1, \ldots, k_j - 1$. For $j = 1, 2, \ldots, n$ and $i = 1, 2, \ldots k_j - 1$, and for each $x \in D$, we define $b_i^{(j)}(x)$ to be 1 if and only if $x_j \geq \beta_i^{(j)}$. Let x^* be the resulting binary vector

$$x^* = (b_1^{(1)}(x), \ldots, b_{k_1}^{(1)}(x), \ldots, b_1^{(n)}(x), \ldots, b_{k_n}^{(n)}(x)) \in \{0, 1\}^d,$$

where $d = \sum_{j=1}^n k_j$. The set $D^* = \{x_i^* : 1 \leq i \leq m\}$ is then a binarized version of the set D of observations, and standard LAD techniques can be applied.

There are a number of ways, however, in which the binarization just described could be non-optimal and, usually, some cutpoints can be eliminated; see the approaches taken in [8,9]. In [5], variants on these approaches are discussed, the aim being to find 'robust' cutpoints; that is, cutpoints which define hyperplanes geometrically at least at a certain distance from the data points. Suppose, then, that a (reduced) set $C^{(j)}$ of K_j cutpoints (a subset of the corresponding $\beta_i^{(j)}$) is selected for coordinate j, and suppose the members of $C^{(j)}$ are

$$a_1^{(j)} < a_2^{(j)} < \cdots < a_{K_j}^{(j)}.$$

Let $d = \sum_{j=1}^n K_j$. An element $x \in [0, 1]^n$ will be 'binarized' as $x^* \in \{0, 1\}^d$ where x^* is

$$(b_1^{(1)}(x), \ldots, b_{K_1}^{(1)}(x), \ldots, b_1^{(n)}(x), \ldots, b_{K_n}^{(n)}(x)),$$

where $b_i^{(j)}(x) = 1$ if and only if $x_j \geq a_i^{(j)}$. Let the Boolean literal $u_i^{(j)}$ be given by $\mathbb{I}[x_j \geq a_i^{(j)}]$, where $\mathbb{I}[P]$ has value 1 if P is true and value 0 otherwise. Then a positive pattern is a conjunction of some of the Boolean variables $u_i^{(j)}$. By definition of $u_i^{(j)}$, $u_i^{(j)} = 1$ implies $u_{i'}^{(j)} = 1$ for $i > i'$, and any j. So a typical positive pattern can be written in terms of these Boolean variables as

$$\bigwedge_{j=1}^n u_{r_j}^{(j)} \bar{u}_{s_j}^{(j)},$$

where $s_j > r_j$. (Here, $\wedge$ denotes the Boolean conjunction, the 'and' operator.) Geometrically, this positive pattern is the indicator function of the 'box'

$$[a_{r_1}^{(1)}, a_{s_1}^{(1)}) \times [a_{r_2}^{(2)}, a_{s_2}^{(2)}) \times \cdots \times [a_{r_n}^{(n)}, a_{s_n}^{(n)}).$$

Please cite this article in press as: M. Anthony, J. Ratsaby, A hybrid classifier based on boxes and nearest neighbors, Discrete Applied Mathematics (2014), http://dx.doi.org/10.1016/j.dam.2014.02.018

With this approach, then, the simplest LAD classifier, which corresponds to a disjunctive normal form, is the indicator function of a union of boxes of this type; and all other regions of $[0, 1]^n$ are classified as negative. With the use also of negative patterns, we then have two separate unions of boxes: one labeled as positive, and the other negative. The other regions of $[0, 1]^n$ must also be classified and, as mentioned above, this is often done by a discriminator, the simplest approach being to classify a point as positive if and only if it is covered by at least as many positive patterns as negative patterns. (See [9,8], for instance.) This would, by default, classify as positive a point or a region which is covered by no patterns at all, of either type. Alternatively, such points or regions could be classified randomly as either positive or negative. One issue we seek to address in this paper is whether a more justifiable approach can be taken to classifying such regions.

3. A hybrid classifier based on boxes and distance

3.1. Definition of the classifiers

The classifiers we study here are in many ways similar to those that result, as just described, from the use of positive and negative patterns in the logical analysis of numerical data. (But their use is not confined to LAD.) However, we combine the use of boxes with the use of a nearest-neighbor paradigm. Explicitly, if a point of $[0, 1]^n$ is not in the union of 'positive' boxes (the region covered by positive patterns) or in the union of 'negative' boxes, then it is not simply classified as positive, nor is it randomly classified. Instead, we take into account the distance of the point from these two unions of boxes. If it is 'closer' to the positive boxes than the negative ones, we classify it as positive. We now describe the classifiers.

For each j between 1 and n, suppose there is a set $C^{(j)} = \{a_1^{(j)}, a_2^{(j)}, \ldots, a_{K_j}^{(j)}\} \subseteq [0, 1]$. We call these the cutpoint sets, by analogy with the LAD terminology. An open box (defined with respect to the cutpoint sets) is a set of the form

$$(a_{r_1}^{(1)}, a_{s_1}^{(1)}) \times (a_{r_2}^{(2)}, a_{s_2}^{(2)}) \times \cdots \times (a_{r_n}^{(n)}, a_{s_n}^{(n)}),$$

where $0 \leq r_j < s_j \leq K_j+1$ (and where $a_0^{(j)}$ is interpreted as 0 and $a_{K_j+1}^{(j)}$ as 1). Note that the 'sides' of the box in each dimension, j, are defined by two cutpoints from $C^{(j)}$ (or the end-points 0, 1). These cutpoints need not be consecutive cutpoints. The cutpoint sets $C^{(j)}$ define $\prod_{j=1}^n \binom{K_j+2}{2}$ open boxes. Now take S_+ and S_- to be unions of some such boxes, in such a way that S_+ and S_- are disjoint. The boxes in S_+ are *positive* (labeled 1) and those in S_- negative (labeled 0); and, generally, there are unlabeled boxes, not in S_+ or S_-.

For a point $x \in X = [0, 1]^n$ let

$$\|x\| := \|x\|_\infty = \max_{1 \leq j \leq n} |x_j|$$

denote the max-norm of x. For two points $x, x' \in X$ the distance between them is $\|x - x'\|$ and for a set $S \subseteq X$ we define the distance from x to S to be $\text{dist}(x, S) = \inf_{x' \in S} \|x - x'\|$. Clearly, for $x \in S$, $\text{dist}(x, S) = 0$.

Given the pair S_+ and S_- of unions of open boxes, denote their closures by $\overline{S_+}$ and $\overline{S_-}$. (So, these are just the same unions of boxes, but with the boundaries included.) We define

$$f_+(x) = \text{dist}\left(x, \overline{S_+}\right), \qquad f_-(x) = \text{dist}\left(x, \overline{S_-}\right) \tag{1}$$

and we let

$$f(x) = \frac{f_-(x) - f_+(x)}{2}.$$

So each pair (S_+, S_-) has a unique f associated with it.

Our classifiers will be the binary functions of the form

$$h(x) = \text{sgn}(f(x)),$$

where $\text{sgn}(z) = 1$ if $z \geq 0$ and $\text{sgn}(z) = 0$ if $z < 0$. So, if F is the set of all real-valued functions of the form

$$f = \frac{f_- - f_+}{2},$$

where f_+ and f_- correspond to unions of boxes S_+ and S_-, then the set of classifiers is

$$H = \{\text{sgn}(f) \mid f \in F\}.$$

Classification in which each category or class is a union of boxes is a long-studied and natural method for pattern classification. It is central, for instance, as noted above, to the methods used for logical analysis of data (see, for example [8,9,14,5]) and has been more widely studied as a geometrical classifier (see [10], for instance). More recently, unions of boxes have been used in combination with a nearest-neighbor (or proximity) paradigm for multi-category classification [13]. (In Felici et al. [13], the taxi-cab or d_1 metric is used rather than the max-norm used here, though we can easily adapt the analysis to apply to the d_1 metric.) Felici et al. [13] use an agglomerative box-clustering method to produce a set of candidate classifiers and then select from these one that is, in a sense they define, optimal. They provide some experimental evidence that this approach works. This paper provides theoretical justification for similar types of classifier.

Please cite this article in press as: M. Anthony, J. Ratsaby, A hybrid classifier based on boxes and nearest neighbors, Discrete Applied Mathematics (2014), http://dx.doi.org/10.1016/j.dam.2014.02.018

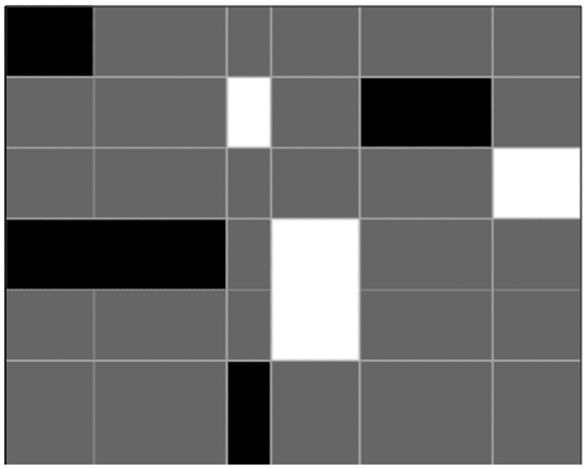

Fig. 1. 'Before classification': the labeled boxes.

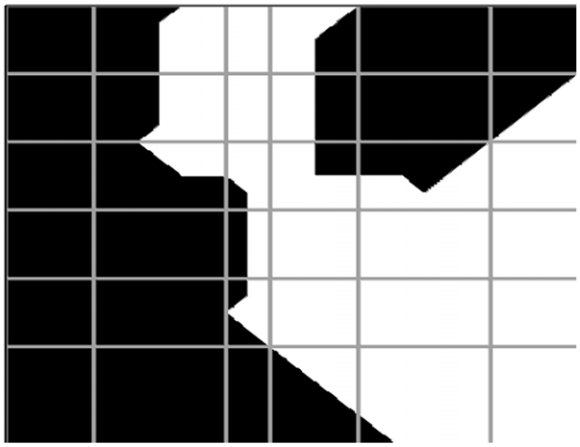

Fig. 2. 'After classification': the classification of the whole domain.

3.2. Example

Fig. 1 shows a 2-dimensional example. We have five cutpoints in each dimension. The white boxes form S_+ and the black boxes form S_-. The gray region is the region which will be classified, in our method, using the distance to the boxes of each type (the 'nearest-neighbor' paradigm). When the whole domain is classified in the way described above, we obtain the partition indicated in Fig. 2: the white region is labeled 1 and the black region 0.

3.3. Rationale

We are interested in this particular kind of classifier for several reasons. A special case of it corresponds quite naturally to a very simple and intuitive learning algorithm. Assume that the cutpoints have the property that we can find boxes defined by them, each of which contains only positive or only negative observations from the known data set. (The standard LAD algorithms for cutpoint selection guarantee this.) Then we could simply take S_+ to be the union of all boxes containing positive observations and S_- the union of those containing negative observations. Any other point x of the domain is then classified according to whether it is closer to the positive boxes or the negative ones.

Furthermore, these classifiers can be used in conjunction with LAD-type methods. One could run an LAD-type algorithm to produce positive and negative patterns. Each pattern corresponds to a box. Let R_+ be the union of the boxes defined by positive patterns and R_- the union of the boxes defined by negative patterns. The fact that some points (and indeed some box regions) could be covered by both positive and negative patterns means that R_+ and R_- need not be disjoint. The intersection $R_+ \cap R_-$ would itself be a union of boxes, and these could be classified, as in standard LAD procedures, using a discriminant. This would assign a box in the intersection a positive classification if and only if the number of positive patterns covering it (that is, boxes of R_+ containing it) is at least the number of negative patterns covering it (boxes of R_- containing it). The classification of these (sub-) boxes would then be resolved, and we could then form S_+ and S_- as the unions of the boxes

Please cite this article in press as: M. Anthony, J. Ratsaby, A hybrid classifier based on boxes and nearest neighbors, Discrete Applied Mathematics (2014), http://dx.doi.org/10.1016/j.dam.2014.02.018

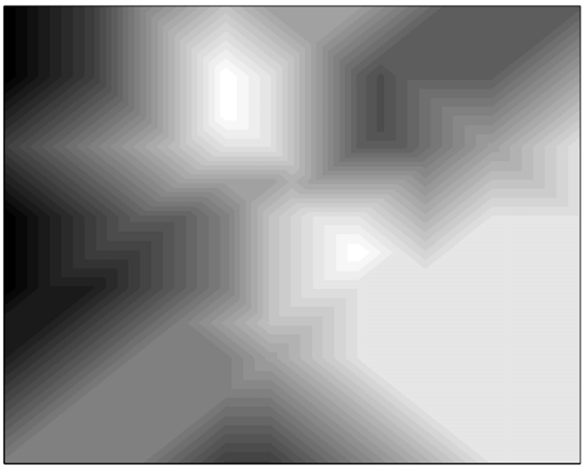

Fig. 3. Contour plot of the underlying real-valued function f. Very dark regions have lowest values of f and very light highest.

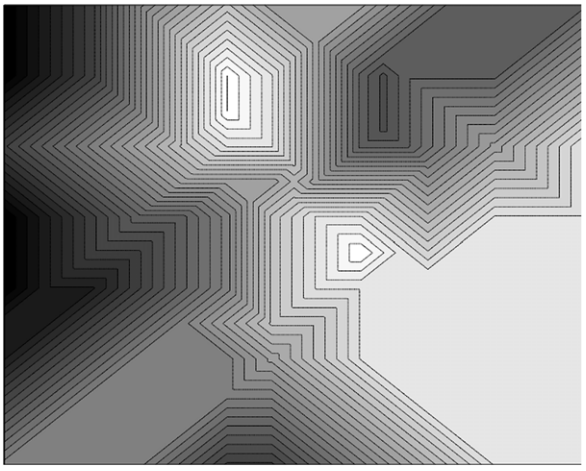

Fig. 4. A contour plot of f, with contour lines indicated.

now labeled 1 and 0, respectively. Then, any point not falling into $S_+ \cup S_-$ (that is, any point not covered by any pattern, positive or negative) is not simply classified as positive by default, but is classified according to whether it is closer to the positive region or the negative region.

Another attractive feature of the classifier produced is that it has a representation which, unlike 'black-box' classification schemes (for instance, based on neural networks), can be described and understood: there are box-shaped regions where we assert a known classification, and the classification anywhere else is determined by an arguably fairly sensible nearest-neighbor approach.

It is also useful that there is an underlying *real-valued* function f. This, as we will see, is useful in analyzing the performance of the classifier. Moreover, the *value* of f (not just its sign) has some geometrical significance. In particular, if $f(x)$ is relatively large, it means that x is quite far from boxes of the opposite classification: it is *not* the case that x is very near the boundary of a box which has the opposite classification. If all points of the data set satisfy this, then it means that the classification is, in a sense, 'robust'. (In related work in [5], a similar notion of robustness of the cutpoints for standard LAD methods is investigated and algorithms for selecting robust cutpoints are discussed.) We could interpret the value of the function f as an indicator of how confident we might be about the classification of a point: a point in the domain with a large value of f will be classified as positive, and more 'definitely' so than one with a smaller, but still positive, value of f. We might think that the classification of the first point is more reliable than that of the second, because the large value of f, indicating that the point is far from negative boxes, provides strong justification for a positive classification. For instance, consider again the example we have been studying. A contour plot of the function f is shown in Fig. 3. The darkest regions are those with lowest (that is, most negative) values of f and the lightest are those with highest value of f. The very dark or very light regions are, arguably, those for which we can most confidently classify the points. Fig. 4 has some contour lines indicated.

Please cite this article in press as: M. Anthony, J. Ratsaby, A hybrid classifier based on boxes and nearest neighbors, Discrete Applied Mathematics (2014), http://dx.doi.org/10.1016/j.dam.2014.02.018

4. Predictive performance of the classifier

4.1. Probabilistic modeling of learning

To quantify the performance of a classifier after training, we use a form of the 'probably approximately correct' (or 'PAC') model of computational learning theory (see [3,17,7]). This assumes that we have some training examples $z_i = (x_i, b_i) \in Z = [0, 1]^n \times \{0, 1\}$, each of which has been generated randomly according to some fixed probability measure P on Z. These training examples are, in the LAD terminology, the labeled positive and negative observations we are given at the outset. Then, we can regard a training sample of length m, which is an element of Z^m, as being randomly generated according to the product probability measure P^m. Suppose that F is the set of functions we are using to classify. (So, recall that F is a set of real-valued functions and that the corresponding binary classification functions are the functions $h = \text{sgn}(f)$ for $f \in F$.)

The natural way to measure the predictive accuracy of $f \in F$ in this context is by the probability that the sign of f correctly classifies future randomly drawn examples. We therefore use the following error measure of the classifier $h = \text{sgn}(f)$:

$$\text{er}_P(\text{sgn}(f)) = P\left(\{(x, b) \in Z : \text{sgn}(f(x)) \neq b\}\right).$$

Of course, we do not know this error: we only know how well the classifier performs on the training sample. We could quantify how well $f \in F$ matches the training sample by using the *sample error* of $h = \text{sgn}(f)$ on the training sample $\mathbf{z} = ((x_1, b_1), \ldots, (x_m, b_m))$:

$$\text{er}_{\mathbf{z}}(h) = \frac{1}{m}|\{i : \text{sgn}(f(x_i)) \neq b_i\}|$$

(the proportion of points in the sample incorrectly classified by the sign of f). But we will find it more useful to use a variant of this, involving a 'width' or 'margin' parameter γ. Much emphasis has been placed in practical machine learning techniques, such as Support Vector Machines [12], on 'learning with a large margin'. (See, for instance [16,1,2,15].) Related work involving 'width' (applicable to binary-valued rather than real-valued functions) rather than 'margin' has also been carried out [4]. For $\gamma > 0$, with $h = \text{sgn}(f)$, and $\mathbf{z} = ((x_1, b_1), \ldots, (x_m, b_m))$, we define

$$\text{er}^{\gamma}_{\mathbf{z}}(f) = \frac{1}{m}|\{i : f(x_i)b_i < \gamma\}|.$$

This is the proportion of $z_i = (x_i, b_i)$ in the sample for which *either* $\text{sgn}(f(x_i)) \neq b_i$, *or* $\text{sgn}(f(x_i)) = b_i$ *but* $|f(x_i)| < \gamma$. So it is the fraction of the sample that is either misclassified by the classifier, or is correctly classified but *not definitively so*, in the sense that the value of $f(x_i)$ is *only just* of the right sign (and not correct 'with a margin' of at least γ).

Much effort has gone into obtaining high-probability bounds on $\text{er}_P(h)$ in terms of $\text{er}^{\gamma}_{\mathbf{z}}(f)$; see [1,6,15,2], for instance. A typical result would be of the following form: for all $\delta \in (0, 1)$, with probability at least $1 - \delta$, for all $f \in F$,

$$\text{er}_P(\text{sgn}(f)) < \text{er}^{\gamma}_{\mathbf{z}}(f) + \epsilon(m, \gamma, \delta),$$

where ϵ decreases with m and δ. We obtain a bound of a similar, but slightly different form, in this paper for the set of hybrid classifiers we are considering.

4.2. Covering the set of classifiers

We now consider *covering numbers*, in order to deploy some results on probabilistic learning. Suppose that F is a set of functions from a domain X to some bounded subset Y of $\mathbb{R}$. For a finite subset S of X, the $l_\infty(S)$-norm of f is defined by $\|f\|_{l_\infty(S)} = \max_{x \in S} |f(x)|$. For $\gamma > 0$, a γ-cover of F with respect to the $l_\infty(S)$ norm is a subset $\hat{F}$ of F with the property that for each $f \in F$ there exists $\hat{f} \in \hat{F}$ with the property that for all $x \in S$, $|f(x) - \hat{f}(x)| < \gamma$. The *covering number* $\mathcal{N}(F, \gamma, l_\infty(S))$ is the smallest cardinality of a covering for F with respect to $l_\infty(S)$ and the *uniform covering number* $\mathcal{N}_\infty(F, \gamma, m)$ is the maximum of $\mathcal{N}(F, \gamma, l_\infty(S))$, over all S with $S \subseteq X$ and $|S| = m$.

We will make use of the following result, a slight improvement of a result that follows from one in [1]. Unlike some standard bounds (see [1], for instance), this has a factor of 3 in front of the $\text{er}^{\gamma}_{\mathbf{z}}(f)$, but it involves ϵ rather than ϵ^2 in the negative exponential. This type of bound is therefore potentially more useful when $\text{er}_{\mathbf{z}}(f)$ is small.

Theorem 4.1. *Suppose that F is a set of $[0, 1]$-valued functions defined on a domain X and that P is any probability measure on $Z = X \times \{0, 1\}$. Then, for any $\epsilon \in (0, 1)$, any $\gamma > 0$ and any positive integer m,*

$$P^m\left(\left\{\mathbf{z} \in Z^m : \exists f \in F, \text{er}_P(\text{sgn}(f)) > 3\,\text{er}^{\gamma}_{\mathbf{z}}(f) + \epsilon\right\}\right) \leq 4\mathcal{N}_\infty(F, \gamma/2, 2m)e^{-\epsilon m/4}.$$

Proof. A theorem from [6] states that, for any η, with probability at least $1 - 4\mathcal{N}_\infty(F, \gamma/2, 2m)e^{-\eta^2 m/4}$, for all $f \in F$,

$$\frac{\text{er}_P(\text{sgn}(f)) - \text{er}^{\gamma}_{\mathbf{z}}(f)}{\sqrt{\text{er}_P(\text{sgn}(f))}} \leq \eta.$$

Please cite this article in press as: M. Anthony, J. Ratsaby, A hybrid classifier based on boxes and nearest neighbors, Discrete Applied Mathematics (2014), http://dx.doi.org/10.1016/j.dam.2014.02.018

Let $\eta = \sqrt{\epsilon}$. Fix $f \in F$ and suppose this inequality holds. If we write x for $\sqrt{\mathrm{er}_P(\mathrm{sgn}(f))}$ and if we set $\beta = \mathrm{er}_{\mathbf{z}}^{\gamma}(f)$, then $x^2 \leq \beta + \eta x$. Equivalently, $x^2 - \eta x - \beta \leq 0$. Regarding this as a quadratic inequality in x, we deduce that

$$x \leq \frac{\eta}{2} + \frac{1}{2}\sqrt{\eta^2 + 4\beta}.$$

It follows that

$$\begin{aligned} x^2 &\leq \frac{\eta^2}{4} + \left(\frac{\eta^2}{4} + \beta\right) + \frac{1}{2}\eta\sqrt{\eta^2 + 4\beta} \\ &\leq \frac{\eta^2}{2} + \beta + \frac{1}{2}\sqrt{\eta^2 + 4\beta}\sqrt{\eta^2 + 4\beta} \\ &= \eta^2 + 3\beta. \end{aligned}$$

So, by fixing $\eta = \sqrt{\epsilon}$, with probability at least $1 - 4\mathcal{N}_\infty(F, \gamma/2, 2m)e^{-\eta^2 m/4}$, for all $f \in F$,

$$\mathrm{er}_P(\mathrm{sgn}(f)) \leq 3\,\mathrm{er}_{\mathbf{z}}^{\gamma}(f) + \epsilon.$$

Hence the result follows (on noting that $\eta^2 = \epsilon$). □

One approach to bounding the covering number of a function class F with respect to the $l_\infty(S)$-norm is to construct and bound the size of a covering with respect to the sup-norm $\|f\|_\infty$ on X, defined as $\|f\|_\infty = \sup_{x\in X} |f(x)|$. This clearly also serves as a covering with respect to the $l_\infty(S)$ norm, for any S, since if $\|f - \hat{f}\|_\infty < \gamma$ then, by definition of the sup-norm, $|f(x) - \hat{f}(x)| < \gamma$ for all $x \in X$ (and, therefore, for all $x \in S$ where S is some subset of X). This is the approach we now take.

The following result will be useful to us.

Lemma 4.2. *Suppose f_+ is defined as in Eq. (1) with respect to the set S_+, a union of boxes based on cutpoints $a_i^{(j)}$ (for $1 \leq j \leq n$ and $1 \leq i \leq K_j$). Then, for any $x \in [0,1]^n$, there exists a pair of indices $1 \leq q \leq n$, $1 \leq p \leq K_q$ such that the distance between x and $\overline{S_+}$ satisfies dist $\left(x, \overline{S_+}\right) = \left|x_q - a_p^{(q)}\right|$.*

Proof. We have

$$\mathrm{dist}\left(x, \overline{S_+}\right) = \inf_{x' \in \overline{S_+}} \phi\left(x, x'\right)$$

where, for each fixed x, $\phi(x, x') = \max_{1\leq j\leq n} |x_j - x'_j|$ is continuous in x'. Since the set $\overline{S_+}$ is closed, by the extreme-value theorem $\phi(x, x')$ attains its infimum on $\overline{S_+}$. If $x \in S_+$ then it is attained at $x' = x$. If $x \notin S_+$ then it is attained at some point on the boundary of S_+. This boundary consists of a union of 'sides', each side j being a set of the form

$$V_i^{(j)} = \left\{ z \in [0,1]^n : z_j = a_i^{(j)},\ a_{r_k}^{(k)} \leq z_k \leq a_{s_k}^{(k)},\ k \neq j \right\}.$$

The point z^* closest to x in $V_i^{(j)}$ is then such that $z_j^* = a_i^{(j)}$, and for the other coordinates $k \neq j$, either $z_k^* = x_k$, or z_k^* equals one of the two cutpoint values $a_{r_k}^{(k)}, a_{s_k}^{(k)}$. Thus the distance $\|x - z^*\|$ either equals $\left|x_j - a_i^{(j)}\right|$; or $\left|x_k - a_{r_k}^{(k)}\right|$ or $\left|x_k - a_{s_k}^{(k)}\right|$, for some $k \neq j$. The distance between x and $\overline{S_+}$ equals the minimal distance between x and any of the sides $V_i^{(j)}$. It follows that this distance equals $\left|x_s - a_p^{(q)}\right|$ for some $1 \leq q \leq n$, $1 \leq p \leq K_q$. □

We now bound the covering number of the set of classifiers that derive from a set of boxes with cardinality at most B. Suppose $B \in \mathbb{N}$ and that for each j between 1 and n, $K_j \in \mathbb{N}$. Let $\ell = (K_1, K_2, \ldots, K_n) \in \mathbb{N}^n$ and let $F(\ell, B)$ be the set of all the classifiers obtained as follows: (i) for each j, there is a set $C^{(j)} = \{a_1^{(j)}, a_2^{(j)}, \ldots, a_{K_j}^{(j)}\} \subseteq [0,1]$; and, (ii) the boxes taken to form $S_+ \cup S_-$ are at most B in number, and each of them has the form

$$(a_{r_1}^{(1)}, a_{s_1}^{(1)}) \times (a_{r_2}^{(2)}, a_{s_2}^{(2)}) \times \cdots \times (a_{r_n}^{(n)}, a_{s_n}^{(n)})$$

where $0 \leq r_j < s_j \leq K_j + 1$ (and where $a_0^{(j)}$ is interpreted as 0 and $a_{K_j+1}^{(j)}$ as 1). (Note that we specify here the *numbers* K_j of cutpoints in each dimension, but we do not fix the *sets* of cutpoints. Note also that the boxes need not be disjoint.)

We have the following bound.

Theorem 4.3. *Let $B \in \mathbb{N}$ and $\ell = (K_1, K_2, \ldots, K_n) \in \mathbb{N}^n$. Then, one has the following bound for the uniform covering numbers of $F(\ell, B)$:*

$$\ln \mathcal{N}_\infty(F(\ell, B), \gamma, m) \leq \sum_{j=1}^{n} K_j \ln\left(\frac{3}{\gamma}\right) + 2B\sum_{j=1}^{n} \ln(K_j + 2) + B,$$

for all $m \in \mathbb{N}$ and all $\gamma \in (0,1)$.

Please cite this article in press as: M. Anthony, J. Ratsaby, A hybrid classifier based on boxes and nearest neighbors, Discrete Applied Mathematics (2014), http://dx.doi.org/10.1016/j.dam.2014.02.018

Proof. As indicated in the preceding discussion, we construct a covering of $F(\ell, B)$ with respect to the sup-norm $\|f\|_\infty$ on $F(\ell, B)$. For a given $\gamma \in (0, 1)$, let $N = \lfloor 1/\gamma \rfloor$ and let

$$G_\gamma = \{i\gamma : 0 \le i \le N\} \cup \{1\} = \left\{0, \gamma, 2\gamma, \ldots, \left\lfloor \frac{1}{\gamma} \right\rfloor \gamma, 1\right\} \subseteq [0, 1].$$

Note that $|G_\gamma| \le N + 2 = \lfloor 1/\gamma \rfloor + 2 \le \lfloor 3/\gamma \rfloor$. Let us define the class $\hat{F}(\ell, B)$ of classifiers to be those satisfying: (i) for each j, there are K_j cutpoints in dimension j, and each belongs to G_γ; and, (ii) the boxes in $S_+ \cup S_-$ are at most B in number. Then we claim that $\hat{F}(\ell, B)$ is a γ-covering of $F(\ell, B)$ with respect to the sup-norm.

Given any $f \in F(\ell, B)$ let $C^{(j)}$ be the set of cutpoints $\{a_1^{(j)}, a_2^{(j)}, \ldots, a_{K_j}^{(j)}\}$, for $1 \le j \le n$. By construction, for each $a_i^{(j)}$ there exists a corresponding $\hat{a}_i^{(j)} \in G_\gamma$ that satisfies $\left|a_i^{(j)} - \hat{a}_i^{(j)}\right| \le \gamma$. For each box Q in S_+ of the form

$$Q = (a_{r_1}^{(1)}, a_{s_1}^{(1)}) \times (a_{r_2}^{(2)}, a_{s_2}^{(2)}) \times \cdots \times (a_{r_n}^{(n)}, a_{s_n}^{(n)}),$$

let $\hat{Q}$ be the box

$$\hat{Q} = (\hat{a}_{r_1}^{(1)}, \hat{a}_{s_1}^{(1)}) \times (\hat{a}_{r_2}^{(2)}, \hat{a}_{s_2}^{(2)}) \times \cdots \times (\hat{a}_{r_n}^{(n)}, \hat{a}_{s_n}^{(n)}).$$

Let $\hat{S}_+$ be the union of the boxes $\hat{Q}$ corresponding to the boxes Q forming S_+. In an analogous way, define $\hat{S}_-$. The function class $\hat{F}(\ell, B)$ is precisely the set of all functions $\hat{f}$, defined by

$$\hat{f}(x) = \frac{\hat{f}_-(x) - \hat{f}_+(x)}{2},$$

where

$$\hat{f}_-(x) = \operatorname{dist}(x, \overline{\hat{S}_-}), \qquad \hat{f}_+(x) = \operatorname{dist}(x, \overline{\hat{S}_+}).$$

We now show that $\|f - \hat{f}\|_\infty \le \gamma$.

Fix any $x \in X$. Let us compute the values of $f_+(x)$ and $\hat{f}_+(x)$. From Lemma 4.2, there exist indices r, s such that $f_+(x) = \operatorname{dist}\left(x, \overline{S_+}\right) = \left|x_s - a_r^{(s)}\right|$. Denote by $\hat{a}_r^{(s)}$ the cutpoint in G_γ that satisfies $\left|\hat{a}_r^{(s)} - a_r^{(s)}\right| \le \gamma$. Then we have,

$$\begin{aligned} f_+(x) &= \left|x_s - a_r^{(s)}\right| \\ &\ge \left|x_s - \hat{a}_r^{(s)}\right| - \gamma \\ &\ge \inf\{\|x - z\| : z \in \overline{\hat{S}_+}\} - \gamma \\ &= \operatorname{dist}\left(x, \overline{\hat{S}_+}\right) - \gamma \\ &= \hat{f}_+(x) - \gamma. \end{aligned}$$

Also, from Lemma 4.2, there exist indices p, q such that $\hat{f}_+(x) = \operatorname{dist}\left(x, \overline{\hat{S}_+}\right) = \left|x_p - \hat{a}_q^{(p)}\right|$. Hence we have

$$\begin{aligned} \hat{f}_+(x) &= \left|x_p - \hat{a}_q^{(p)}\right| \\ &\ge \left|x_p - a_q^{(p)}\right| - \gamma \\ &\ge \inf\{\|x - z\| : z \in \overline{S_+}\} - \gamma \\ &= \operatorname{dist}\left(x, \overline{S_+}\right) - \gamma \\ &= f_+(x) - \gamma. \end{aligned}$$

It follows that $\|f_+ - \hat{f}_+\| \le \gamma$. The same argument holds for the pair f_- and $\hat{f}_-$, and so it follows that

$$\begin{aligned} \|f - \hat{f}\|_\infty &= \sup_{x \in X} \left|f(x) - \hat{f}(x)\right| \\ &= \frac{1}{2} \sup_{x \in X} \left|f_-(x) - f_+(x) - \hat{f}_-(x) + \hat{f}_+(x)\right| \\ &\le \frac{1}{2} \sup_{x \in X} \left|f_+(x) - \hat{f}_+(x)\right| + \frac{1}{2} \sup_{x \in X} \left|f_-(x) - \hat{f}_-(x)\right| \\ &\le \gamma. \end{aligned}$$

Thus for each $f \in F(\ell, B)$ there exists $\hat{f} \in \hat{F}(\ell, B)$ such that $\|f - \hat{f}\|_\infty \le \gamma$, and $\hat{F}(\ell, B)$ is therefore a γ-covering of $F(\ell, B)$ in the sup-norm.

Please cite this article in press as: M. Anthony, J. Ratsaby, A hybrid classifier based on boxes and nearest neighbors, Discrete Applied Mathematics (2014), http://dx.doi.org/10.1016/j.dam.2014.02.018

We now bound the cardinality of $\hat{F}(\ell, B)$. Note that since there are K_j cutpoints in each dimension j, and each of these is from G_γ, a set of cardinality at most $\lfloor 3/\gamma \rfloor$, it follows that there are at most $\prod_{j=1}^{n} \binom{\lfloor 3/\gamma \rfloor}{K_j}$ possible ways of choosing the cutpoints for a function in $\hat{F}(\ell, B)$. A box is defined by choosing a pair of cutpoints in each dimension (allowing also for the possibility that one end of the interval defining the box in any given dimension can be 0 or 1). We then choose B boxes, and, next, each box is assigned either a 0 label or a 1 label (that is, it is chosen to be part of $\hat{S}_-$ or $\hat{S}_+$). Thus, we have

$$|\hat{F}(\ell, B)| \leq \prod_{j=1}^{n} \binom{\lfloor 3/\gamma \rfloor}{K_j} \binom{\prod_{j=1}^{n} \binom{K_j+2}{2}}{B} 2^B$$
$$\leq \prod_{j=1}^{n} \left\lfloor \frac{3}{\gamma} \right\rfloor^{K_j} \prod_{j=1}^{n} (K_j+2)^{2B} 2^B.$$

It follows, therefore, that

$$\ln |\hat{F}(\ell, B)| \leq \sum_{j=1}^{n} K_j \ln\left(\frac{3}{\gamma}\right) + 2B \sum_{j=1}^{n} \ln(K_j+2) + B$$

and the result of the theorem follows. □

4.3. A generalization error bound

Theorem 4.1, together with Theorem 4.3, could now be used to bound the generalization error of a classifier when B and $K_1, K_2, \ldots, K_n$ are prescribed in advance. However, the following more useful result does not require these to be known or prescribed.

Theorem 4.4. *Suppose $\delta \in (0, 1)$, and suppose P is any probability measure on $X = [0, 1]^n$. Then, with P^m-probability at least $1 - \delta$, a sample $\mathbf{z}$ is such that:*

- *for all $\gamma \in (0, 1)$;*
- *for all $\ell = (K_1, K_2, \ldots, K_n) \in \mathbb{N}^n$;*
- *for all $B \in \mathbb{N}$;*
- *if $f \in F(\ell, B)$, then*

$$\mathrm{er}_P(\mathrm{sgn}(f)) \leq 3\,\mathrm{er}_{\mathbf{z}}^{\gamma}(f) + \epsilon(m, \gamma, \delta, \ell, B),$$

where $\epsilon(m, \gamma, \delta, \ell, B)$ is

$$\frac{4}{m}\left(\ln\left(\frac{12}{\gamma\delta}\right) + \sum_{j=1}^{n} K_j \ln\left(\frac{24}{\gamma}\right) + 2B + 2B\sum_{j=1}^{n} \ln(K_j+2)\right).$$

Proof. Theorems 4.1 and 4.3 have the following immediate consequence: for $\ell \in \mathbb{N}^n$ and $B \in \mathbb{N}$, with probability at least $1 - \delta$, for all $f \in F(\ell, B)$,

$$\mathrm{er}_P(\mathrm{sgn}(f)) < 3\,\mathrm{er}_{\mathbf{z}}^{\gamma}(f) + \epsilon_1(m, \gamma, \delta, \ell, B)$$

where $\epsilon_1(m, \gamma, \delta, \ell, B)$ is

$$\frac{4}{m}\left(\ln\left(\frac{4}{\delta}\right) + \sum_{j=1}^{n} K_j \ln\left(\frac{6}{\delta}\right) + B + 2B\sum_{j=1}^{n} \ln(K_j+2)\right).$$

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 $f \in F(\ell, B)$ with $\mathrm{er}_P(\mathrm{sgn}(f)) \geq 3\,\mathrm{er}_{\mathbf{z}}^{\alpha_2}(f) + \epsilon_1(m, \alpha_1, \delta, \ell, B)$. Then, as just noted, $P^m(E(\alpha, \alpha, \delta)) \leq \delta$ and, also, if $\alpha_1 \leq \alpha \leq \alpha_2$ and $\delta_1 \leq \delta$, then $E(\alpha_1, \alpha_2, \delta_1) \subseteq E(\alpha, \alpha, \delta)$. It follows, from [6,1], that

$$P^m\left(\bigcup_{\alpha \in (0,1]} E(\alpha/2, \alpha, \delta\alpha/2)\right) \leq \delta.$$

In other words, for fixed ℓ and B, with probability at least $1 - \delta$, *for all* $\gamma \in (0, 1]$, we have

$$\mathrm{er}_P(\mathrm{sgn}(f)) < 3\,\mathrm{er}_{\mathbf{z}}^{\gamma}(f) + \epsilon_2(m, \gamma, \delta, \ell, B),$$

Please cite this article in press as: M. Anthony, J. Ratsaby, A hybrid classifier based on boxes and nearest neighbors, Discrete Applied Mathematics (2014), http://dx.doi.org/10.1016/j.dam.2014.02.018

where $\epsilon_2(m, \gamma, \delta, \ell, B)$ is

$$\frac{4}{m}\left(\ln\left(\frac{4}{\delta}\right)+\sum_{j=1}^{n} K_j \ln\left(\frac{12}{\delta\gamma}\right)+B+2B\sum_{j=1}^{n}\ln(K_j+2)\right).$$

(Note that γ now need not be prescribed in advance.) It now follows that the probability that for *some* ℓ and for *some* B, we have

$$\mathrm{er}_P(\mathrm{sgn}(f)) \geq 3\,\mathrm{er}_{\mathbf{z}}^{\gamma}(f)+\epsilon_2\left(m, \gamma, \frac{\delta}{2^{(B+\sum_{j=1}^{n} K_j)}}, \ell, B\right)$$

for *some* $\gamma \in (0, 1)$ is at most

$$\begin{aligned}\sum_{B,K_1,\ldots,K_n=1}^{\infty}\frac{\delta}{2^{(B+\sum_{j=1}^{n} K_j)}} &= \sum_{B=1}^{\infty}\frac{\delta}{2^B}\sum_{K_1,\ldots,K_n=1}^{\infty}\prod_{j=1}^{n}\frac{1}{2^{K_j}}\\ &= \sum_{B=1}^{\infty}\frac{\delta}{2^B}\prod_{j=1}^{n}\sum_{K_j=1}^{\infty}\frac{1}{2^{K_j}}\\ &= \sum_{B=1}^{\infty}\frac{\delta}{2^B}\prod_{j=1}^{n}1=\delta.\end{aligned}$$

The result now follows. □

For any classifier of the type considered, there will be some maximal value of γ such that $\mathrm{er}_{\mathbf{z}}^{\gamma}(f)=0$. We call this value of γ the *width* of f on $\mathbf{z}$. This terminology is motivated by the earlier observation that the value of $f(x)$ measures the distance from x to the nearest box with the opposite classification. (The term 'margin' might be more standard in the general context of using real-valued functions for classification, but 'width' seems more geometrically appropriate here.) Theorem 4.4 does not specify γ in advance, so we have the following immediate corollary.

Theorem 4.5. *With the same notation as above, with P^m-probability at least $1-\delta$, a sample $\mathbf{z}$ is such that for any $f \in F$, $\mathrm{er}_P(\mathrm{sgn}(f))$ is at most*

$$\frac{4}{m}\left(\ln\left(\frac{12}{\gamma(f,\mathbf{z})\delta}\right)+\sum_{j=1}^{n} K_j \ln\left(\frac{24}{\gamma(f,\mathbf{z})}\right)+2B(f)+B(f)\sum_{j=1}^{n}\ln(K_j+1)\right),$$

where $\gamma(f,\mathbf{z})$ is the width of f on $\mathbf{z}$, and f involves $B(f)$ boxes, defined with respect to some set of K_j cutpoints in dimension j (for $1 \leq j \leq n$).

We could also use Theorem 4.4 as a guide to 'model selection'. The theorem states that, with probability at least $1-\delta$,

$$\mathrm{er}_P(\mathrm{sgn}(f)) < E(m, \gamma, \delta, \ell, B) = 3\,\mathrm{er}_{\mathbf{z}}^{\gamma}(f)+\epsilon(m, \gamma, \delta, \ell, B).$$

For fixed m and δ, $\epsilon(m, \gamma, \delta, \ell, B)$ decreases as γ increases, and $\mathrm{er}_{\mathbf{z}}^{\gamma}(f)$ increases as γ increases. Therefore $E(m, \gamma, \delta, \ell, B)$ is the sum of two quantities, one increasing and the other decreasing as γ increases, and there is hence a trade-off between the two quantities. Clearly, also, the parameters ℓ and B can be varied. This motivates the use of a learning algorithm that returns a classifier which minimizes the combination $E(m, \gamma, \delta, \ell, B)$. The (high-probability) generalization error bound for such an algorithm takes the form

$$\mathrm{er}_P(\mathrm{sgn}(f)) \leq \inf_{\gamma,\ell,B}\inf\{3\,\mathrm{er}_{\mathbf{z}}^{\gamma}(f)+\epsilon(m, \gamma, \delta, \ell, B) : f \in F(\ell, B)\}.$$

How to develop an algorithmic procedure for realizing such a learning algorithm is an interesting question for further work; but, certainly, what this suggests is that, given a choice of possible classifiers, it might be sensible to select from among those the one that minimizes the right-hand side of this bound.

5. Conclusions

This paper has studied the generalization ability of a classifier that is a hybrid between two approaches: classical LAD (or unions of boxes) and nearest-neighbors. In using real-valued functions as a key tool in the classification, we can attach some measure of how 'definitive' the classification is. This is potentially of some practical use, but it also enables us to develop generalization error bounds of a special type, which depend on a measure of the classifier's robustness, which we term its 'width'.

Acknowledgments

This work was supported in part by the IST Programme of the European Community, under the PASCAL2 Network of Excellence, IST-2007-216886. Joel Ratsaby acknowledges the support of the Department of Computer Science at UCL.

Please cite this article in press as: M. Anthony, J. Ratsaby, A hybrid classifier based on boxes and nearest neighbors, Discrete Applied Mathematics (2014), http://dx.doi.org/10.1016/j.dam.2014.02.018

References

[1] M. Anthony, P.L. Bartlett, Neural Network Learning: Theoretical Foundations, Cambridge University Press, 1999.
[2] M. Anthony, P.L. Bartlett, Function learning from interpolation, Combin. Probab. Comput. 9 (2000) 213–225.
[3] M. Anthony, N. Biggs, Computational Learning Theory, in: Cambridge Tracts in Theoretical Computer Science, vol. 30, Cambridge University Press, 1992 Reprinted 1997.
[4] M. Anthony, J. Ratsaby, Maximal width learning of binary functions, Theoret. Comput. Sci. 411 (2010) 138–147.
[5] M. Anthony, J. Ratsaby, Robust cutpoints in the logical analysis of numerical data, Discrete Appl. Math. 160 (2012) 355–364.
[6] 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. Inform. Theory 44 (2) (1998) 525–536.
[7] A. Blumer, A. Ehrenfeucht, D. Haussler, M.K. Warmuth, Learnability and the Vapnik–Chervonenkis dimension, J. ACM 36 (4) (1989) 929–965.
[8] E. Boros, P.L. Hammer, T. Ibaraki, A. Kogan, Logical analysis of numerical data, Math. Program. 79 (1997) 163–190.
[9] E. Boros, P.L. Hammer, T. Ibaraki, A. Kogan, E. Mayoraz, Ilya Muchnik, An implementation of logical analysis of data, IEEE Trans. Knowl. Data Eng. 12 (2) (2000) 292–306.
[10] N.H. Bshouty, P.W. Goldberg, S.A. Goldman, H.D. Mathias, Exact learning of discretized geometric concepts, SIAM J. Comput. 28 (2) (1998) 674–699.
[11] Y. Crama, P.L. Hammer, T. Ibaraki, Cause-effect relationships and partially defined boolean functions, Ann. Oper. Res. 16 (1988) 299–325.
[12] N. Cristianini, J. Shawe-Taylor, An Introduction to Support Vector Machines and Other Kernel-Based Learning Methods, Cambridge University Press, Cambridge, 2000.
[13] G. Felici, B. Simeone, V. Spinelli, Classification techniques and error control in logic mining, in: Data Mining 2010, in: Annals of Information Systems, vol. 8, 2010, pp. 99–119.
[14] P.L. Hammer, Y. Liu, S. Szedmák, B. Simeone, Saturated systems of homogeneous boxes and the logical analysis of numerical data, Discrete Appl. Math. 144 (1–2) (2004) 103–109.
[15] J. Shawe-Taylor, P.L. Bartlett, R.C. Williamson, M. Anthony, Structural risk minimization over data-dependent hierarchies, IEEE Trans. Inform. Theory 44 (5) (1996) 1926–1940.
[16] A.J. Smola, P.L. Bartlett, B. Scholkopf, D. Schuurmans, Advances in Large-Margin Classifiers, in: Neural Information Processing, MIT Press, 2000.
[17] V.N. Vapnik, Statistical Learning Theory, Wiley, 1998.

Please cite this article in press as: M. Anthony, J. Ratsaby, A hybrid classifier based on boxes and nearest neighbors, Discrete Applied Mathematics (2014), http://dx.doi.org/10.1016/j.dam.2014.02.018