PROBABILISTIC ESTIMATION OF VAPNIK-CHERVONENKIS DIMENSION

Przemysław Klęsk
Department of Methods of Artificial Intelligence and Applied Mathematics, West Pomeranian University of Technology ul. Żolnierska 49, Szczecin, Poland

Keywords: Statistical learning theory, Machine-learning, Vapnik-Chervonenkis dimension, Binary classification.

Abstract: We present an idea of probabilistic estimation of Vapnik-Chervonenkis dimension given a set of indicator functions. The idea is embedded in two algorithms we propose — named A and A'. Both algorithms are based on an approach that can be described as *expand or divide and conquer*. Also, algorithms are parametrized by probabilistic constraints expressed in a form of (ε,δ)-precision. The precision implies how often and by how much the estimate can deviate from the true VC-dimension. Analysis of convergence and computational complexity for proposed algorithms is also presented.

1 INTRODUCTION

Vapnik-Chervonenkis dimension is an important notion within Statistical Learning Theory (Vapnik and Chervonenkis, 1968; Vapnik and Chervonenkis, 1989; Vapnik, 1995; Vapnik, 1998). Many bounds on generalization or sample complexity are based on it.

Recently, several other measures of functions sets capacity (richness) have been under study. Particularly, of great interest are *covering numbers* (Bartlett et al., 1997; Anthony and Bartlett, 2009). In many cases covering numbers can lead to tighter bounds (on generalization or sample complexity) than pessimistic bounds based on VC-dimension. However, the constructive derivation of covering numbers itself is usually a challenge. One has to suitably take advantage of some properties of given set of functions or of the learning algorithm and discover how they translate onto a cover. One of such attractive results is e.g. a result from (Zhang, 2002) related to regularization. Qualitatively, it states that for sets of functions linear in parameters and under a L_q-regularization (general $q = 1,2,\ldots$) the bound on covering number scales only linearly with the dimension of input domain. This allows to learn and generalize well with a sample complexity logarithmic in the number of attributes. On the other hand, there exist results where the property used for the derivation of covering numbers is actually the known VC-dimension of some set of functions (Anthony and Bartlett, 2009), which again proves its usefulness.

Known are some sets of functions for which the exact value of VC-dimension has been established by suitable combinatorial or geometric proofs (often very complex). Here are some examples. For polynomials defined over $\mathbb{R}^d$ of degree at most n, the VC-dim is $\binom{n+d}{d}$, see e.g. (Anthony and Bartlett, 2009). For hyperplanes in $\mathbb{R}^d$ (which can be bases for multilayer perceptrons) the VC-dim is $d+1$ (Vapnik, 1998). For rectangles in $\mathbb{R}^d$ the VC-dim is $2d$ (Cherkassky and Mulier, 1998). For spheres in $\mathbb{R}^d$ (which can be bases of RBF neural networks) the VC-dim is $d+1$ (Cherkassky and Mulier, 1998). As regards linear combinations of bases as above the VC-dim can typically be bounded by the number of bases times the VC-dim of a single base (Anthony and Bartlett, 2009, p. 154), this fact however requires usually a careful analysis.

Also, some analysis has been done in the subject of computational complexity for the VC-dimension. In particular, in (Papadimitriou and Yannakakis, 1996) authors take up the following problem ,,*given a set of functions F and a natural number k, is VC-dim$(F) \geq k$?"*, i.e. one asks about a lower bound of VC-dimension. And the problem is proved to be logNP-complete.

Our motivation for this paper is to introduce an idea for algorithms, which given an arbitrary set of functions (plus a learning algorithm) would be able to estimate its VC-dimension with an imposed *probabilistic* accuracy. Such algorithms, if sufficiently suc-

Klesk P..
PROBABILISTIC ESTIMATION OF VAPNIK-CHERVONENKIS DIMENSION.
DOI: 10.5220/0003721702620270
In *Proceedings of the 4th International Conference on Agents and Artificial Intelligence* (ICAART-2012), pages 262-270
ISBN: 978-989-8425-95-9

cessful, could potentially replace the need for complex proofs establishing the exact value of the VC-dimension.

2 NOTATION, NOTIONS, TOOLS

We restrict considerations to the binary classification learning problems.

Let F denote the set of indicator functions[1], which we have at disposal for learning. Let L denote the learning algorithm we use to choose a single function from F. This happens via the *sample error minimization*[2] principle.

Let P denote the unknown joint probability distribution defined over $\mathbf{Z} = \mathbf{X} \times Y$ from which training pairs $\mathbf{z} = (\mathbf{x}, y)$ are drawn, where in general $\mathbf{x} \in \mathbb{R}^d$ are input points and $y \in \{0,1\}$ are corresponding class labels. By $\mathbf{z} = \{\mathbf{z}_1, \mathbf{z}_2, \ldots, \mathbf{z}_m\}$ we shall denote the whole training sample of size m drawn from the product distribution P^m in a i.i.d. manner[3].

For any fixed function $f \in F$ the true generalization error with respect to P is typically calculated as

$$\mathrm{er}_P(f) = \int_{\mathbf{Z}} l_f(\mathbf{z}) dP(\mathbf{z}), \tag{1}$$

where l_f is the following loss function

$$l_f(\mathbf{z}) = l_f\big((\mathbf{x}, y)\big) = \begin{cases} 0, & \text{for } f(x) = y; \\ 1, & \text{for } f(x) \neq y. \end{cases} \tag{2}$$

Therefore $\mathrm{er}_P(f)$ expresses the probability of misclassification of $(\mathbf{x}, y)$ drawn randomly from P. Since P is unknown the learning algorithm L can only try to minimize the frequency of misclassification on the observed sample i.e.:

$$\widehat{\mathrm{er}}_{\mathbf{z}}(f) = \frac{1}{m} \sum_{i=1}^{m} l_f(\mathbf{z}_i). \tag{3}$$

Let the solution-function of L be denoted by $\widehat{f}$.

We now briefly remind some notions introduced by Vapnik. Let $l_F = \{l_f : f \in F\}$ denote the set of loss functions generated by F. Consider the following set

$$\big\{ \big(l_f(\mathbf{z}_1), \ldots, l_f(\mathbf{z}_m)\big) : f \in F \big\}. \tag{4}$$

It contains all distinguishable functions in l_F restricted to the fixed sample $\mathbf{z}_1, \ldots, \mathbf{z}_m$. Throughout the paper we shall denote (4) by $(l_F)_{|\mathbf{z}_1, \ldots, \mathbf{z}_m}$.

[1] $\{0,1\}$-valued functions.

[2] Alternatively also called *empirical risk minimization*

[3] Independent, identically distributed. This means P^m is unknown but fixed.

Using a natural correspondence between indicator functions and dichotomies of a set, Vapnik introduces the notion of *shattering*. We say that l_F shatters a sample $\mathbf{z}_1, \ldots, \mathbf{z}_m$ if all its dichotomies can be generated using functions from F, equivalently this means that the number of distinguishable functions is $\#(l_F)_{|\mathbf{z}_1, \ldots, \mathbf{z}_m} = 2^m$. The Vapnik-Chervonenkis dimension for l_F (or equivalently for F) is equal to the size of some largest sample that can be shattered.

It will be helpful to remind three more quantities:

- *Vapnik-Chervonenkis entropy*

$$H^F(m) = \int_{\mathbf{Z}^m} \ln \#(l_F)_{|\mathbf{z}_1, \ldots, \mathbf{z}_m} dP^m(\mathbf{z}_1, \ldots, \mathbf{z}_m), \tag{5}$$

  which is an expectation of the logarithm of the number of distinguishable functions;

- *annealed entropy*

$$H^F_{\mathrm{ann}}(m) = \ln \int_{\mathbf{Z}^m} \#(l_F)_{|\mathbf{z}_1, \ldots, \mathbf{z}_m} dP^m(\mathbf{z}_1, \ldots, \mathbf{z}_m), \tag{6}$$

  which is a logarithm of expected number of distinguishable functions;

- *growth function*

$$G^F(m) = \sup_{\mathbf{z}_1, \ldots, \mathbf{z}_m} \#(l_F)_{|\mathbf{z}_1, \ldots, \mathbf{z}_m}, \tag{7}$$

  which is supremum number of distinguishable functions.

Known is the connection $H^F(m) \leq H^F_{\mathrm{ann}}(m) \leq \ln G^F(m)$, where the first inequality is due to Jensen inequality. Known also is the fact, that VC-dim(F) is equal to such an argument of G^F after which it stops growing exponentially.

As a tool, throughout the paper, we shall extensively take advantage of one-sided Chernoff inequalities (Hellman and Raviv, 1970; Schmidt et al., 1995), which we now write down the following way

$$p - \nu_m \leq \sqrt{\frac{-\ln \delta}{2m}}, \tag{8}$$

$$\nu_m - p \leq \sqrt{\frac{-\ln \delta}{2m}}, \tag{9}$$

where p is a probability of some event (that will be of interest for us) and ν_m is its frequency observed in m independent trials. Each inequality holds true with probability[4] at least $1 - \delta$.

Also, in several places we are going to take advantage of Iverson notation $[s]$, which returns 1 if the statement s is true and 0 otherwise (Graham et al., 2002).

[4] The $1 - \delta$ is an outer probability calculated with respect to probabilistic space defined over all random experiments consisting of m independent trials.

3 PROBABILISTIC SHATTERING

In this section we introduce several new notions, which can be regarded as probabilistic versions of selected notions reminded in the previous section. The new notions are suitable for our purposes and give a high-level intuition on algorithms we are about to propose.

3.1 Distribution Dependence — Two Conceptual Scenarios

We start with the following remark. It is the fact that: shattering, growth function and VC-dimension are *distribution-independent* notions. For our purposes it will be convenient though to define notions that are distribution-dependent, because we are going to carry out probabilistic estimations. All the notions shall therefore refer to P or P^m. Two conceptual scenarios are possible here.

I. In this scenario we think of P as it was originally defined — i.e. the joint probability distribution defined over $\mathbf{X} \times Y$ describing the specific learning problem. And therefore we should treat all new notions as *distribution-dependent* counterparts of classical Vapnik's notions.

II. In this scenario we conceptually replace P by the *uniform distribution*. By doing so we separate ourselves from the specific problem. For this purpose, we only need to assume a boundedness of $\mathbf{X}$. The P will still explicitly appear in the notions and formulas. But, we can then agree (as a form of convention) to look at the notions as *distribution-independent* or at least 'original problem'-distribution-independent, since the uniformness does not favor any samples.

The reader can therefore treat further considerations in either context — of scenario I or II. In both scenarios we shall assume that we can freely and numerously redraw samples from P.

3.2 New Notions

Definition 1. *We say that $\mu^F(m)$ is a **shatterability measure** with respect to the probability distribution P^m, and is calculated as follows*

$$\mu^F(m) = \int_{\mathbf{Z}^m} [\#(l_F)_{|\mathbf{z}_1,\ldots,\mathbf{z}_m} = 2^m] dP^m(\mathbf{z}_1,\ldots,\mathbf{z}_m). \tag{10}$$

Intuitively the *shatterability measure* expresses how frequently one 'comes across' samples drawn from P^m which can be shattered. We suggest to think of shatterability measure in conjunction with the growth function $G^F(m)$, see (7). Imagine some method trying to discover the argument $\mathbf{z}_1,\ldots,\mathbf{z}_m$ in P^m for which the supremum is attained. Of course for strictness, we must remind that firstly the definition of $G^F(m)$ is distribution-independent and secondly even if it was distribution-dependent then the supremum could be attained on sets of measure zero. Nevertheless, the intuition that the smaller $\mu^F(m)$ the more difficult it is to indicate the supremum represented by $G^F(m)$ is true. In particular if $G^F(m) < 2^m$ then certainly $\mu^F(m) = 0$.

Definition 2. *We say that a set of indicator functions F is an m-shatterer with respect to P^m (or: shatters some samples of size m drawn from P^m) if $\mu^F(m) > 0$.*

Definition 3. *We say that a set of indicator functions F is not an m-shatterer with respect to P^m everywhere, if the two conditions are met:*

1. $\mu^F(m) = 0$,
2. $\nexists\, \mathbf{z}_1,\ldots,\mathbf{z}_m$ *such that* $\#(l_F)_{|\mathbf{z}_1,\ldots,\mathbf{z}_m} = 2^m$.

Definition 4. *We say that a set of indicator functions F is not an m-shatterer with respect to P^m almost everywhere, if the two conditions are met:*

1. $\exists\, \mathbf{z}_1,\ldots,\mathbf{z}_m$ *such that* $\#(l_F)_{|\mathbf{z}_1,\ldots,\mathbf{z}_m} = 2^m$,
2. $\mu^F(m) = 0$.

The complementary definitions above follow from the arguments discussed earlier, and the *almost everywhere* condition takes into account that the case where 2^m dichotomies are feasible but for sets (samples) of measure zero.

3.3 Probabilistic Estimation of VC-dimension — Sketch of Idea

We now sketch an idea according to which the algorithms to be presented later shall work.

Suppose that for given sample of size m we execute multiple times (say n times) an experiment consisting of drawing a sample $\mathbf{z}_1,\ldots,\mathbf{z}_m$ from P^m and checking exhaustively if all its dichotomies are feasible, i.e. checking if $\#(l_F)_{|\mathbf{z}_1,\ldots,\mathbf{z}_m} = 2^m$. If for any experiment this is true, then we can stop (before n is reached), since certainly $\mathrm{VCdim}(F) \geq m$ and we can try to increase the sample size. If this event did not occur in any experiment, then by means of Chernoff inequality we have that with probability at least $1-\delta$:

$$\mu^F(m) \leq 0 + \sqrt{\frac{-\ln\delta}{2n}}. \tag{11}$$

We write down 0 explicitly on purpose — it is the observed frequency of the event 'all dichotomies are

feasible on random sample'. In that case we shall decrease the sample size. We would also like to introduce a probabilistic precision parameter for the algorithm. We name it (ε,δ)-precision, $0<\varepsilon,\delta<1$. If we insert $\varepsilon := \sqrt{-\ln\delta/(2n)}$, it follows that the needed number of experiments is $n = \lceil -\ln\delta/(2\varepsilon^2)\rceil$.

Now, by analogy to the definition 4, we introduce the following definition.

Definition 5. *We say that a set of indicator functions F is not a (m,ε,δ)-shatterer with respect to P^m if with probability at least $1-\delta$:*

$$\mu^F(m) \leq \varepsilon.$$

In simple words we say (with an imposed probabilistic precision) that F does not shatter samples of size m, if the probability that 2^m dichotomies on a random sample are feasible is suitably small.

Now, we define the *probabilistic VC-dimension*.

Definition 6. *We say that the probabilistic (ε,δ)-VC-dimension for the set F equals m, we write*

$$VCdim_{\varepsilon,\delta}(F) = m,$$

if there exists a sample of size m that can be shattered by F and simultaneously F is not a $(m+1,\varepsilon,\delta)$-shatterer.

4 ALGORITHM A

The algorithm A, we are about to propose, returns the probabilistic dimension $\mathrm{VCdim}_{\varepsilon,\delta}(F)$. This value is an estimate of the true VC-dimension.

First, we present an auxiliary algorithm called B, which will be invoked by the main algorithm A in a loop. The algorithm B works as the checker of feasibility of all dichotomies given a fixed sample, accordingly to the sketch from the section 3.3. The algorithm returns 1 when all dichotomies are feasible and 0 otherwise.

$B(F;\mathbf{z}_1,\ldots,\mathbf{z}_m)$

1. For all $(t_1,\ldots,t_m)\in\{0,1\}^m$:
   1.1. Create a temporary training sample $S = (x_1,t_1),\ldots,(x_m,t_m)$ and execute learning algorithm L on it, which yields $\widehat{f}$.
   1.2. If $\widehat{\mathrm{er}}_S(\widehat{f}) > 0$ return 0.
2. Return 1.

Figure 1: Auxiliary algorithm B.

We now present the algorithm A which works with an imposed (ε,δ)-precision, see the Fig. 2. As arguments for A, apart from F we also enlist P, with solely such an intention that we will be able to draw multiple samples from it, nothing more (since P in general can be unknown, recall scenario I).

$A_{\varepsilon,\delta}(F,P)$

1. Set $m_L := 1$, $m_U := \infty$, $m := m_L$.
2. Repeat while $m_U - m_L > 1$:
   2.1. Set $s := 0$.
   2.2. Repeat $n = \lceil -\ln\delta/(2\varepsilon^2)\rceil$ times:
      2.2.1 Draw a sample $\mathbf{z}_1,\ldots,\mathbf{z}_m$ from P^m.
      2.2.2 If $B(F;\mathbf{z}_1,\ldots,\mathbf{z}_m) = 1$ then set $s := 1$ and jump out of the loop 2.2.
   2.3 If $m_U = \infty$:
      2.3.1 If $s = 1$ then set $m_L := 2m$, $m := m_L$.
      2.3.2 Else set $m_L := 1/2m$, $m_U := m$, $m := (m_L + m_U)/2$.
   2.4 Else
      2.4.1 If $s = 1$ then set $m_L := m$, $m := (m_L + m_U)/2$.
      2.4.2 Else set $m_U := m$, $m := (m_L + m_U)/2$.
3. Return $\lfloor m_L \rfloor$.

Figure 2: Algorithm A.

The algorithm uses an approach that could be described as *expand or divide and conquer*. At the start we set the lower bound m_L and the current sample size m to 1, whereas we set the upper bound m_U to infinity. At first, as the algorithm progresses and all dichotomies prove feasible (s flag equals 1), the tested sample sizes are doubled (step 2.3.1.). Let us call it the *expand-phase*. When a moment is reached such that all dichotomies are not feasible despite n trials, the algorithm suitably sets m_L and m_U (no longer infinite) and puts the next sample size m to be tested in the middle of m_L and m_U (step 2.3.2.). This moment starts the *divide-phase*. Since then, all next executions of the main loop (step 2) make the algorithm enter step 2.4. and suitably narrow down the interval $[m_L, m_U)$ until the stop condition is reached.

The form of the return value $\lfloor m_L \rfloor$ requires a short explanation. The floor function is meant to handle the special case when after the first iteration of the main loop (step 2.) the s flag is already equal 0. Then halfening (step 2.3.2.) causes m_L to be $1/2$, and since the stop condition is reached we want to correct this value to 0.

5 CONVERGENCE AND COMPUTATIONAL COMPLEXITY ANALYSIS

We will show that it is convenient to analyze con-

vergence of the algorithm A in terms of shatterability measures for given problem.

5.1 Sequence of Shatterability Measures — General Observations

Consider the sequence of shatterability measures along growing sample size:

$$\mu^F(1), \mu^F(2), \ldots.$$

A moment of thought leads to the following observation.

Lemma 1. *The sequence $\mu^F(1), \mu^F(2), \ldots$ is non-increasing.*

Proof. By independence and Fubini's theorem we have that:

$$\begin{aligned}
&\mu^F(m+1) \\
&= \int_{\mathbf{Z}^{m+1}} [\#(l_F)_{|\mathbf{z}_1,\ldots,\mathbf{z}_{m+1}} = 2^{m+1}] dP^{m+1}(\mathbf{z}_1,\ldots,\mathbf{z}_{m+1}) \\
&= \int_{\mathbf{Z}} \int_{\mathbf{Z}^m} [\#(l_F)_{|\mathbf{z}_1,\ldots,\mathbf{z}_{m+1}} = 2^{m+1}] dP^m(\mathbf{z}_1,\ldots,\mathbf{z}_m) dP(\mathbf{z}_{m+1}) \\
&\leq \int_{\mathbf{Z}} \int_{\mathbf{Z}^m} [\#(l_F)_{|\mathbf{z}_1,\ldots,\mathbf{z}_m} = 2^m] dP^m(\mathbf{z}_1,\ldots,\mathbf{z}_m) dP(\mathbf{z}_{m+1}) \\
&\leq \mu^F(m) \int_{\mathbf{Z}} dP(\mathbf{z}_{m+1}).
\end{aligned}$$

□

Please note that in the second equality-pass any $\mathbf{z}_i$ can be taken outside the inner integral, not necessarily $\mathbf{z}_{m+1}$, and the rest of the proof is still valid.

A second obvious observation is that $\mu^F(m) = 0$ for all $m > \mathrm{VCdim}(F)$. This follows from the definition of VC-dimension.

A more interesting fact is that there exist sets of functions F and distributions P for which the sequence complies with the following pattern: $(1,\ldots,1,0,\ldots)$. It means the sequence consists solely of starting ones and after some point zeros take place. Consider e.g. hyperplanes on a plane. Clearly any single point or two points can be shattered by a hyperplane. Any three points can also be shattered provided that they do not lie in the same line. This is called a "general position", see e.g. (Anthony and Bartlett, 2009, Theorem 3.1), (Wenocur and Dudley, 1981). But even so, the situation of three points lying in the same line is of probability measure zero in continuous spaces. Therefore the sequence for that case would be $(1,1,1,0,\ldots)$. On the other hand it is possible to indicate certain sets F and distributions P for which the sequence that does not consist solely of ones and zeros. As an example see the Fig. 3. It illustrates a set of functions defined over a plane with the decision boundary in the shape of 'U' letter. Suppose 'U' is of *fixed* width and height and it can be shifted *only* along horizontal axes. As the figure shows there exist samples of size $m = 1$ (with positive probability measure) for which only 1 dichotomy is feasible. Also, there exist such samples (also with positive probability measure) for which 2 dichotomies are feasible. The same is true for the case of $m = 2$. Therefore, the corresponding shatterability measures must be fractions.

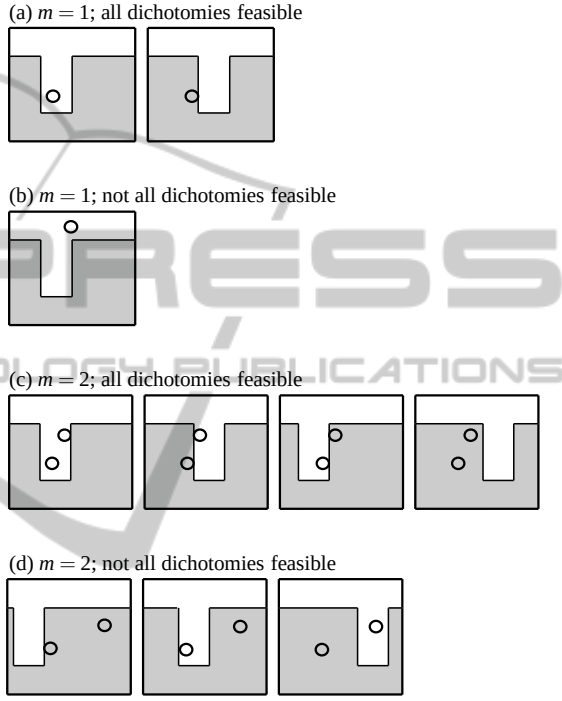

Figure 3: Set of functions with horizontally shifting 'U'-shaped decision boundary of fixed width. Illustration of feasibility of all dichotomies for different samples.

From now one, for shortness we will denote the sequence by $\mu_1, \mu_2, \ldots$.

5.2 Results Distribution and Convergence for Algorithm A

As one may note, the result of algorithm A being $\mathrm{VCdim}_{\varepsilon,\delta}(F)$ cannot be an overestimation of the true $\mathrm{VCdim}(F)$, but it might be its underestimation. In this section we analyze how often this underestimation takes place and in effect we derive the probability distribution defined over the results to which A can converge. The analysis is carried out in terms of the sequence $\mu_1, \mu_2, \ldots$.

Let $p(h)$ denote the probability that A returns $\mathrm{VCdim}_{\varepsilon,\delta}(F) = h$ and let us start by taking a closer look at small cases. For $h = 0$ we have

$$p(0) = (1-\mu_1)^n, \tag{12}$$

since it requires that in all $n = \lceil -\ln\delta/(2\varepsilon^2)\rceil$ independent trials the event opposite to feasibility of all dichotomies occurs (in n times the algorithm B returned 0). For $h = 1$ we have

$$p(1) = \big(1-(1-\mu_1)^n\big)(1-\mu_2)^n. \tag{13}$$

The first factor arises as a complement of $p(0)$ — the algorithm discovered that for some sample of size $m = 1$ all dichotomies were feasible, but it failed to discover such property for $m = 2$, hence the second factor. The cases of $h = 2,3$ reveal more of the *expand or divide and conquer* approach:

$$p(2) = \big(1-(1-\mu_1)^n\big)\big(1-(1-\mu_2)^n\big)(1-\mu_4)^n (1-\mu_3)^n. \tag{14}$$

$$p(3) = \big(1-(1-\mu_1)^n\big)\big(1-(1-\mu_2)^n\big)(1-\mu_4)^n \big(1-(1-\mu_3)^n\big), \tag{15}$$

After the algorithm failed to discover feasibility of all dichotomies for $m = 4$, it had to make a jump backwards to check the case of $m = 3$. We now move to a bigger case example of $h = 21$ which illustrates well forward and backward jumps during the *divide phase* in a chronological order (see indices of μ).

$$p(21) = \big(1-(1-\mu_1)^n\big)\big(1-(1-\mu_2)^n\big)\big(1-(1-\mu_4)^n\big) \big(1-(1-\mu_8)^n\big)\big(1-(1-\mu_{16})^n\big)(1-\mu_{32})^n (1-\mu_{24})^n\big(1-(1-\mu_{20})^n\big)(1-\mu_{22})^n\big(1-(1-\mu_{21})^n\big). \tag{16}$$

A careful analysis allows to find a regular formula for the whole distribution. We state it as the following theorem.

Theorem 1. *Suppose $\mu_1,\mu_2,\ldots$ is the sequence of shatterability measures for given set of functions F and distribution P. Let $q = \lfloor \log_2 h \rfloor$ and let $(h_q, h_{q-1},\ldots,h_0)_2$ denote a binary representation for each $h > 0$. Then, the probability distribution of results to which algorithm A may converge is:*

$$\begin{aligned} p(0) &= (1-\mu_1)^n, \\ p(1) &= \big(1-(1-\mu_1)^n\big)(1-\mu_2)^n. \\ p(h) &= \prod_{k=0}^{q}\big(1-(1-\mu_{2^k})^n\big)(1-\mu_{2^{q+1}})^n \\ &\quad \cdot \prod_{k=0}^{q-1}\big(h_{q-k-1} + (-1)^{h_{q-k-1}}(1-\mu_{i(h,k)})^n\big), \end{aligned} \tag{17}$$

for $h \geq 2$, where

$$i(h,k) = \frac{1}{2}(2^{q+1}+2^q) + \sum_{j=1}^{k}(-1)^{1-h_{q-j}}\cdot 2^{q-j-1}. \tag{18}$$

Sketch of proof. Note that during the *expand phase* the algorithm performs $\lfloor \log_2 h \rfloor + 2$ iterations (which is $q+2$) and this is represented in $p(h)$ by the product $\prod_{k=0}^{q}\big(1-(1-\mu_{2^k})^n\big)(1-\mu_{2^{q+1}})^n$. In this product all but last factors must be of form $1-(1-\mu_{2^k})^n$, since the algorithm discovered that some sample of size 2^k can be shattered, whereas the last factor must be of form $(1-\mu_{2^{q+1}})^n$, since in n trials samples of size 2^{q+1} failed to be shattered. In the *divide phase* the algorithm performs $\log_2(2^{q+1}-2^q) = q$ iterations, this is represented by the remaining product. The $i(h,k)$ function handles suitably successive indices visited by the algorithm and it is easy to check that these indices are determined by the $q-1$ least significant bits in the binary representation $(h_q, h_{q-1},\ldots,h_0)_2$. These bits determine also whether the factor should be of form $(1-\mu_{i(h,k)})^n$ or $1-(1-\mu_{i(h,k)})^n$. □

The following statements are direct consequences of $p(h)$ distribution.

Corollary 1. *Suppose that $VCdim(F) = h^*$ and suppose the sequence of shatterability measures for given F and P consists solely of ones and zeros. Then distribution of results is $p(h^*) = 1$ and $p(h) = 0$ for all $h \neq h^*$. Therefore, for any $0 < \varepsilon,\delta < 1$ we have that $A_{\varepsilon,\delta}(F,P) = h^*$.*

This states that the algorithm A always converges to the true Vapnik-Chervonenkis dimension if the sequence of shatterability measures does not contain fractions.

Corollary 2. *Suppose that $VCdim(F) = h^*$ and suppose the sequence of shatterability measures contains fractions. Then the expected result is $\mathbb{E}A_{\varepsilon,\delta}(F,P) < h^*$, where expectation is taken over infinite number of runs of algorithm A for given problem.*

This states that the algorithm A underestimates the true Vapnik-Chervonenkis dimension if the sequence of shatterability measures does contain fractions.

5.3 Computational Complexity

It is easy to see that the number of iterations of the main loop in algorithm A (step 2.) is logarithmic as a function of the true $\mathrm{VCdim}(F) = h^*$. The number of iterations is at most $2\log_2 h^* + 2$. Recall that there are $q+2$ iterations needed by the *expand phase* and q iterations by the *divide phase*. Unfortunately the most heavy step is the execution of the algorithm B (step 2.2.2.), since it is an exhaustive check of feasibility of all dichotomies. Therefore if we consider the computational complexity as a function of ε,δ,h^* then the pessimistic number of iterations

$$\sum_{\substack{\text{visited indices}\\ \text{of } \mu_i}} n2^i \leq n\sum_{i=1}^{2h^*} 2^i = O\Big(n(2^{h^*+1}-1)\Big), \quad (19)$$

which is exponential in h^*. This is a consequence of the fact the B is an exact algorithm.

In the next section we propose a new algorithm named A'. It is very similar to A but uses an auxiliary algorithm B' being a softened probabilistic version of B. This leads to a constant (at most) complexity of the step 2.1.2. and in effect logarithmic complexity of the whole algorithm.

6 ALGORITHM A'

First, we formulate a probabilistic auxiliary algorithm B'. For a fixed sample $\mathbf{z}_1,\ldots,\mathbf{z}_m$ consider the following quantity: $\eta^F(\mathbf{z}_1,\ldots,\mathbf{z}_m)$ defined as the probability that a random dichotomy drawn from the uniform distribution (defined over $\{0,1\}^m$) is feasible by some function in F on $\mathbf{z}_1,\ldots,\mathbf{z}_m$:

$$\eta^F(\mathbf{z}_1,\ldots,\mathbf{z}_m) = \quad (20)$$

$$\sum_{i=0}^{2^m-1} \frac{1}{2^m}[\exists\, f\in F \text{ realizing dichotomy } (i_{m-1},\ldots,i_0)_2 \text{ on } \mathbf{z}_1,\ldots,\mathbf{z}_m], \quad (21)$$

where $(i_{m-1},\ldots,i_0)_2$ is a binary representation of i.

We shall introduce an additional (ε,δ)-precision. Suppose we would like to have $B'_{\varepsilon,\delta}(F;\mathbf{z}_1,\ldots,\mathbf{z}_m) = 0$ if an unfeasible dichotomy occurred, and to have $B'_{\varepsilon,\delta}(F;\mathbf{z}_1,\ldots,\mathbf{z}_m) = 1$ if with probability at least $1-\delta$

$$\eta^F(\mathbf{z}_1,\ldots,\mathbf{z}_m) \geq 1-\varepsilon \quad (22)$$

holds true.

The algorithm B' is presented in the Fig. 4.

$B'_{\varepsilon,\delta}(F;\mathbf{z}_1,\ldots,\mathbf{z}_m)$

1. Repeat $N = -\ln\delta/(2\varepsilon^2)$ times:
   1.1. Draw a random dichotomy $(t_1,\ldots,t_m)$ from a uniform distribution.
   1.2. Create a temporary training sample $S = (x_1,t_1),\ldots,(x_m,t_m)$ and execute learning algorithm L on it, which yields $\widehat{f}$.
   1.3. If $\widehat{\mathrm{er}}_S(\widehat{f}) > 0$ return 0.
2. Return 1.

Figure 4: Auxiliary algorithm B'.

We now present the algorithm A'. Since the inner auxiliary algorithm was probabilistically softened, the algorithm A' requires now four precision parameters $\varepsilon_1,\delta_1,\varepsilon_2,\delta_2$, see the Fig. 5.

$A'_{\varepsilon_1,\delta_1,\varepsilon_2,\delta_2}(F,P)$

1. Set $m_L := 1$, $m_U := \infty$, $m := m_L$.
2. Repeat while $m_U - m_L > 1$:
   2.1. Set $s := 0$.
   2.2. Repeat $n = \lceil -\ln\delta_1/(2\varepsilon_1^2)\rceil$ times:
      2.2.1 Draw a sample $\mathbf{z}_1,\ldots,\mathbf{z}_m$ from P^m.
      2.2.2 If $B'_{\varepsilon_2,\delta_2}(F;\mathbf{z}_1,\ldots,\mathbf{z}_m) = 1$ then set $s := 1$ and jump out of the loop 2.2.
   2.3 If $m_U = \infty$:
      2.3.1 If $s = 1$ then set $m_L := 2m$, $m := m_L$.
      2.3.2 Else set $m_L := 1/2m$, $m_U := m$, $m := (m_L + m_U)/2$.
   2.4 Else
      2.4.1 If $s = 1$ then set $m_L := m$, $m := (m_L + m_U)/2$.
      2.4.2 Else set $m_U := m$, $m := (m_L + m_U)/2$.
3. Return $\lfloor m_L \rfloor$.

Figure 5: Algorithm A'.

The result of A' is quantity compliant with the following definition (and is an estimation of the true VC-dimension).

Definition 7. *We say that the probabilistic $(\varepsilon_1,\delta_1,\varepsilon_2,\delta_2)$-VC-dimension for the set F equals m, we write*

$$VCdim_{\varepsilon_1,\delta_1,\varepsilon_2,\delta_2}(F) = m,$$

if there exists a sample of size $\mathbf{z}_1,\ldots,\mathbf{z}_m$ such that with probability at least $1-\delta_2$

$$\eta^F(\mathbf{z}_1,\ldots,\mathbf{z}_m) \geq 1-\varepsilon_2 \quad (23)$$

and with probability at least $1-\delta_1$

$$\mu^F(m+1) \leq \varepsilon_1. \quad (24)$$

Putting it in simpler wording, the probabilistic $(\varepsilon_1,\delta_1,\varepsilon_2,\delta_2)$-VC-dimension is m if we can indicate a sample of size m for which with high probability all dichotomies are feasible, and simultaneously with high probability we cannot indicate such sample of size $m+1$. Obviously, both probability parameters refer strictly to quantities μ and η, which one should be aware of. They are related to different probabilistic spaces. The probability $1-\delta_1$ and μ quantities refer to the probabilistic space with P distribution, whereas the probability $1-\delta_2$ and η quantities refer to the probabilistic space describing feasibility of random dichotomies drawn uniformly from $\{0,1\}^m$ for some fixed sample $\mathbf{z}_1,\ldots,\mathbf{z}_m$.

Please note that, in contrast to the algorithm A, the result of A' can be (with small probability) both underestimation and overestimation of the true VCdim(F).

It is worth remarking that the algorithm B' is of constant complexity $O(N)$ where $N = -\ln\delta_2/(2\varepsilon_2^2)$. Therefore, it is easy to see that the complexity of the A' algorithm is

$$O\left(\frac{-\ln\delta_1}{2\varepsilon_1^2}\frac{-\ln\delta_2}{2\varepsilon_2^2}\log_2 h^*\right). \tag{25}$$

6.1 Notes on Distribution of Results for Algorithm A'

Having in mind the theorem 1 which describes the probability distribution $p(h)$ of results to which the algorithm A may converge, we can try to do a similar analysis for the A' algorithm. The main difference now is that A' can overestimate the true VC-dimension. This happens when for some sample drawn in the step 2.2.1. some dichotomies are not feasible, but B' fails to discover it in its N trials. In other words, apart from quantities $\mu^F(m)$ the involvement of $\eta^F(\mathbf{z}_1,\ldots,\mathbf{z}_m)$ must be taken into account.

Consider the following expectation

$$\begin{aligned}\alpha_m = &\int_{\mathbf{Z}^m}\Big([\#(l_F)_{|\mathbf{z}_1,\ldots,\mathbf{z}_m} = 2^m] \\ &+ [\#(l_F)_{|\mathbf{z}_1,\ldots,\mathbf{z}_m} < 2^m]\eta^F(\mathbf{z}_1,\ldots,\mathbf{z}_m)^N\Big) \\ &dP^m(\mathbf{z}_1,\ldots,\mathbf{z}_m) \\ = &\mu^F(m) \\ &+ \int_{\mathbf{Z}^m}[\#(l_F)_{|\mathbf{z}_1,\ldots,\mathbf{z}_m} < 2^m]\eta^F(\mathbf{z}_1,\ldots,\mathbf{z}_m)^N \\ &dP^m(\mathbf{z}_1,\ldots,\mathbf{z}_m).\end{aligned} \tag{26}$$

It describes (in an average case) the probability of an event of interest, i.e. : that either a randomly drawn sample of size m can be shattered (first summand) or it cannot be shattered, but this fact was not discovered in N trials (second summand). Therefore, to explicitly write down the theoretical probability distribution for results of A' it is sufficient to insert into (17) quantities α_i in the place of μ_i.

7 SUMMARY AND FUTURE RESEARCH

In the paper we propose a general idea for probabilistic estimation of the VC-dimension for an arbitrary set of indicator functions. The idea required suitable definitions of several notions and quantities which can be regarded as probabilistic counterparts of some traditional notions defined by Vapnik.

The main idea is based on an approach we call *expand or divide and conquer* and is represented by two algorithms A and A' that we propose. The analysis of computational complexity shows that A' requires only logarithmic time with respect to the true VC-dimension it tries to discover. This time scales also with imposed precision parameters: $n = -\ln\delta_1/(2\varepsilon_1)^2$, $N = -\ln\delta_2/(2\varepsilon_2)^2$, and their scaling influence on the time is $O(n \cdot N)$.

We are aware that the presented part of research constitutes only the theoretical part. Certainly, practical applications of the idea may still require a thorough experimental research first, possibly some refinements in algorithms, in order to be successful. In the future, we plan to carry out the following experimentally-oriented studies on the idea:

1. executions of A and A' on sets of functions with simple geometrical bases (hyperplanes, spheres, rectangles etc.),
2. tests for linear combinations of bases,
3. tests for sets of functions with regularization,
4. tests on convergence and performance,
5. registering histograms of experimental distributions of results to see how heavy are the tails (i.e. how often under/overestimations of the true VC-dimension occur),
6. discovering 'good' settings for precision parameters for given conditions of experiment,
7. tests for sets of functions for which the true VC-dimension is unknown.

Results of these studies ought to form a separate publication.

ACKNOWLEDGEMENTS

This work has been financed by the Polish Government, Ministry of Science and Higher Education from the sources for science within years 2010–2012. Research project no.: N N516 424938.

REFERENCES

Anthony, M. and Bartlett, P. (2009). *Neural Network Learning: Theoretical Foundations*. Cambridge University Press, Cambridge, UK.

Bartlett, P., Kulkarni, S., and Posner, S. (1997). Covering numbers for real-valued function classes. *IEEE Transactions on Information Theory*, 47:1721–1724.

Cherkassky, V. and Mulier, F. (1998). *Learning from data*. John Wiley & Sons, inc.

Graham, R., Knuth, D., and Patashik, O. (2002). *Concrete Mathematics. A foundation for Computer Science*. Wydawnictwo Naukowe PWN SA, Warsaw, Poland.

Hellman, M. and Raviv, J. (1970). Probability of error, equivocation and the chernoff bound. *IEEE Transactions on Information Theory*, IT-16(4):368–372.

Papadimitriou, C. and Yannakakis, M. (1996). On limited nondeterminism and the complexity of the V-C dimension. *Journal of Computer and System Sciences*, 53:161–170.

Schmidt, J., Siegel, A., and Srinivasan, A. (1995). Chernoff-hoeffding bounds for applications with limited independence. *SIAM Journal on Discrete Mathematics*, 8(2):223–250.

Vapnik, V. (1995). *The Nature of Statistical Learning Theory*. Springer Verlag, New York.

Vapnik, V. (1998). *Statistical Learning Theory: Inference from Small Samples*. Wiley, New York.

Vapnik, V. and Chervonenkis, A. (1968). On the uniform convergence of relative frequencies of events to their probabilities. *Dokl. Aka. Nauk*, 181.

Vapnik, V. and Chervonenkis, A. (1989). The necessary and sufficient conditions for the consistency of the method of empirical risk minimization. *Yearbook of the Academy of Sciences of the USSR on Recognition, Classification and Forecasting*, 2:217–249.

Wenocur, R. and Dudley, R. (1981). Some special Vapnik-Chervonenkis classes. *Discrete Mathematics*, 33:313–318.

Zhang, T. (2002). Covering number bounds of certain regularized linear function classes. *Journal of Machine Learning Research*, 2:527–550.