The Optimal PAC Bound for Intersection-Closed Concept Classes

Malte Darnstädt[1]

Ruhr-Universität Bochum, Universitätsstraße 150, 44801 Bochum, Germany

Abstract

Thirty years after the introduction of Valiant's PAC-learning framework in 1984, proving sharp sample complexity bounds in the standard, realizable framework is still an open problem. This is true for general concept classes but also for most natural families of classes. In this letter we will give sharp bounds on the sample complexity of PAC-learning intersection-closed classes. Our result settles an open problem posed by Auer and Ortner and supports a conjecture by Warmuth about the true sample complexity and the optimal PAC-learning algorithm for general classes.

Furthermore this letter demonstrates a useful application of the disagreement coefficient—a complexity measure developed for agnostic learning by Giné and Koltchinskii based on the work of Alexander and, independently, by Hanneke—in the realizable PAC-learning framework.

Keywords: computational complexity, PAC-learning, sample complexity, intersection-closed classes, disagreement coefficient

1. Introduction

Since Valiant introduced the PAC-learning framework in [1], it is an open problem to give sharps bounds on the number of labeled examples necessary to successfully learn in the realizable setting (readers unfamiliar with this framework may want to read Section 2 first, which gives a succinct definition of PAC-learning). The best known worst case lower bound, which was proven by Ehrenfeucht, Haussler, Kearns and Valiant in [2], is

$$\Omega\left(\frac{d+\log(1/\delta)}{\varepsilon}\right)$$

Email address: malte.darnstaedt@rub.de (Malte Darnstädt)

[1]Tel.: +49 234 32 23209

This work was supported by the bilateral Research Support Programme between Germany (DAAD 50751924) and Hungary (MÖB 14440).

where d is the VC-dimension of the concept class, ε is the accuracy and δ is the confidence parameter. On the other hand, the best matching upper bound, which was proven by Blumer, Ehrenfeucht, Haussler and Warmuth [3], is

$$O\left(\frac{d\log(1/\varepsilon)+\log(1/\delta)}{\varepsilon}\right) \ .$$

Obviously, these bounds leave a gap of $\log(1/\varepsilon)$.

Warmuth conjectured in [4] that the factor of $\log(1/\varepsilon)$ in the upper bound can be overcome. He furthermore conjectured that this can be achieved by the *one-inclusion graph algorithm*, which—in the case of intersection-closed concept classes—collapses to the *closure algorithm*, whose output is the smallest consistent hypothesis. For some special intersection-closed classes, which possess an additional combinatorial property, Auer and Ortner provided a proof of the conjecture in [5]. We will show that Warmuth's conjecture is indeed true for *all* intersection-closed classes.

We will prove our main result by employing a technique which involves the so-called *disagreement coefficient*. The disagreement coefficient was first used in learning theory, without calling it by this name, by Giné and Koltchinskii in a paper on agnostic learning in 2006 [6], which was based on the work of Alexander [7], but it was independently discovered by Hanneke in 2007 [8] to analyze the label complexity of agnostic active learning. While most applications of the disagreement coefficient are in the field of agnostic active learning, we are aware of two exceptions where it was used in the standard, i.e., realizable and passive setting: one section of Hanneke's Ph.D. thesis [9, page 50ff] and an article by D., Simon and Szörenyi [10]. We will combine the ideas from these two works and improve on Hanneke's results in Section 4.

2. Preliminaries

2.1. Notation and basic definitions

We introduce some notation and remind the reader of the definitions of PAC-learning [1] and the VC-dimension [11]:

For any set A, let 2^A denote the power set of A. Let X be a nonempty set, called the *domain*, let D be a distribution over X and, for any (measurable) $S \subseteq X$, let $D(S)$ denote the probability mass of S under D. A *concept class*[2] is a set $\mathcal{C} \subseteq 2^X$. The elements of $\mathcal{C}$ are called *concepts* and we identify these sets with their corresponding indicator functions. The *error* of concept c under target concept t and distribution D is defined as $err_{t,D}(c) := \Pr_{x\sim D}\big(c(x) \neq t(x)\big)$.

For any set $S \subseteq X$ and any $t \in \mathcal{C}$, we define the *sample* $S_t := \{(x, t(x)) \,|\, x \in S\}$, i.e. the elements of S paired with labels from $\{0, 1\}$ induced by concept t.[3]

[2] In this letter we do not consider hypothesis classes that are different from the concept class. Our results can easily be extended to that case.

[3] We often draw S i.i.d. according to D, so in fact $S \in X^m$ for some $m > 0$, but we always ignore the order and multiplicities in S and it is easier to think of it as a set.

The set $V(S_t) := \{c \in \mathcal{C} \,|\, c(x) = t(x) \text{ for all } x \in S\}$ is called the *version space* and, for any $V \subseteq \mathcal{C}$, $DIS(V) := \{x \in X \,|\, \exists c_1, c_2 \in V : c_1(x) \neq c_2(x)\}$ is called the *disagreement region of* V. Note that for any $c \in V$, the disagreement region is the union of all subsets from X where some concept from V disagrees with c, i.e. $DIS(V) = \bigcup_{c' \in V} \{x \in X \,|\, c'(x) \neq c(x)\}$. In the case $t \in V$, we see that $DIS(V)$ is the union of all sets where some concept from V errs.

A *PAC-learner* L is an algorithm that receives a sample as its input and outputs a concept, such that for all $t \in \mathcal{C}$, all distributions D over X and all $\varepsilon, \delta > 0$ there exists an m_0, such that for all $m \geq m_0$ it holds that

$$\Pr_{S \sim D^m} \left(err_{t,D}(L(S_t)) \leq \varepsilon \right) \geq 1 - \delta \ .^4$$

We call $L(S_t)$ the *hypothesis* of the learner. The *sample complexity of class* $\mathcal{C}$ is defined as the smallest possible m_0, as a function in ε and δ, for the best PAC-learner of $\mathcal{C}$.

We say that a set $S \subseteq X$ is *shattered* by class $\mathcal{C}$ if $\{c \cap S \,|\, c \in \mathcal{C}\} = 2^S$. The *VC-dimension* of $\mathcal{C}$, denoted by $VC(\mathcal{C})$, is defined as the size of the largest set shattered by $\mathcal{C}$ (or infinite if no such set exists).

2.2. Intersection-closed classes

A class $\mathcal{C}$ is called *intersection-closed* if for all $c_1, c_2 \in \mathcal{C}$ it also holds that $c_1 \cap c_2 \in \mathcal{C}$. For example, the class of Boolean monomials over n variables and the class of axis-aligned boxes in $\mathbb{R}^n$ are intersection-closed.

We will assume in the following that intersection-closed classes contain *all* intersections, i.e. that $\cap_{c \in \mathcal{C}'} c \in \mathcal{C}$ holds for any $\mathcal{C}' \subseteq \mathcal{C}$. As Auer and Ortner have shown in [5], adding the missing intersections does not increase the VC-dimension of $\mathcal{C}$.

We briefly note that all results that hold for intersection-closed classes can also be applied to classes that are 'union-closed' by taking the complement of each concept over X.

3. Disagreement coefficients

Hanneke's original definition of the disagreement coefficient is as follows:

Definition 1 (Hanneke [8]). For $r > 0$ and $t \in \mathcal{C}$, let $B(t, r)$ be the ball of radius r around concept t, i.e. $B(t, r) := \{c \in \mathcal{C} \,|\, err_{t,D}(c) \leq r\}$. Then *Hanneke's disagreement coefficient* is defined as

$$\theta_H(\mathcal{C} \,|\, t, D) := \sup_{r > 0} \frac{D(DIS(B(t, r)))}{r} \ .$$

Taking the supremum over all $t \in \mathcal{C}$ and all distributions D over X yields the distribution independent disagreement coefficient $\theta_H(\mathcal{C}) := \sup_{t,D} \theta_H(\mathcal{C} \,|\, t, D)$.

[4] While Valiant's original definition of PAC-learning also demands a polynomial running time in $1/\varepsilon$ and $1/\delta$, we will ignore computational issues in this letter.

D., Simon and Szörenyi considered the following variant, which will turn out to be more useful in realizable PAC-learning:

Definition 2 (D., Simon and Szörenyi [10]). For $t \in \mathcal{C}$, any $S \subseteq X$ and any distribution D over X, the *realizable disagreement coefficient* is defined as

$$\theta_R(\mathcal{C} \,|\, t, S, D) := \frac{D(DIS(V(S_t)))}{\sup_{c \in V(S_t)} err_{t,D}(c)} \ .$$

For the degenerate case that both denominator and numerator are zero, we define $\theta_R(\mathcal{C} \,|\, t, S, D) := 1$. Similar as before, let $\theta_R(\mathcal{C} \,|\, t, D) := \sup_{S \subseteq X} \theta_R(\mathcal{C} \,|\, t, S, D)$ and $\theta_R(\mathcal{C}) := \sup_{t,D} \theta_R(\mathcal{C} \,|\, t, D)$.

As noted in [10], setting $r^* := \sup_{c \in V(S_t)} err_{t,D}(c)$ immediately shows that $\theta_R(\mathcal{C} \,|\, t, S, D) = D(DIS(V(S_t)))/r^* \leq D(DIS(B(t, r^*)))/r^* \leq \theta_H(\mathcal{C} \,|\, t, D)$. The following two lemmas demonstrate that the realizable disagreement coefficient can be much smaller than Hanneke's:

Lemma 3. *For any class $\mathcal{C}$ over X it holds that $VC(\mathcal{C}) \leq \theta_H(\mathcal{C}) \leq |X|$.*

Proof. The first inequality is trivial for classes of VC-dimension zero, since θ_H is non-negative. Now let $S \subseteq X$ be a set of size $d \geq 1$ shattered by $\mathcal{C}$ and let D be the uniform distribution over S. Then $\theta_H(\mathcal{C} \,|\, t, D) \geq d$ holds for all $t \in \mathcal{C}$: let $r^* = 1/d$, so that $S \subseteq DIS(B(t, r^*))$. Since $D(S) = 1$, it follows that $\theta_H(\mathcal{C} \,|\, t, D) \geq 1/(1/d) = d$.

The second inequality is trivial for infinite X. Note that $\theta_H(\mathcal{C}) \leq |X|$ holds for finite X, because $DIS(B(t, r)) \subseteq \{x \in X \,|\, D(\{x\}) \leq r\}$. □

Lemma 4. *Let $\mathcal{C}$ be a concept class, D a distribution and S_t a sample. If a concept $\tilde{c} \in V(S_t)$ exists, that errs on the whole disagreement region almost surely under D, i.e., $D(DIS(V(S_t)) \setminus \{x \in X \,|\, t(x) \neq \tilde{c}(x)\}) = 0$, then $\theta_R(\mathcal{C} \,|\, t, S, D) = 1$.*

Proof. From $\sup_{c \in V(S_t)} err_{t,D}(c) \geq err_{t,D}(\tilde{c}) \geq D(DIS(V(S_t)))$ follows that $\theta_R(\mathcal{C} \,|\, t, S, D) \leq 1$. On the other hand, the error of a concept from the version space can never be larger than the probability mass of the disagreement region. Thus it holds that $\theta_R(\mathcal{C} \,|\, t, S, D) \geq 1$. □

Example 5. Let $X \neq \emptyset$ be a finite set and let $\mathcal{C} = 2^X$. From Lemma 3 and 4 follows $\theta_H(\mathcal{C}) = |X| \geq 1 = \theta_R(\mathcal{C})$.

4. PAC bounds

While the main application of θ_H is in the field of agnostic active learning, Hanneke also proved the following theorem for the realizable framework:

Theorem 6 (Hanneke [9]). *Let $\mathcal{C}$ be a concept class of VC-dimension d, D a distribution over X, $S \sim D^m$ a random sample of size m and $t \in \mathcal{C}$. Then for any $\delta > 0$ it holds with probability $\geq 1 - \delta$ for all $h \in V(S_t)$*

$$err_{t,D}(h) \leq \frac{24}{m}\left(d \log(880\, \theta_H(\mathcal{C}|t, D)) + \log \frac{12}{\delta}\right) \ .$$

Thus the sample complexity is upper bounded by $O\left(\frac{1}{\varepsilon}(d \log \theta_H(\mathcal{C}) + \log \frac{1}{\delta})\right)$.

Combining Hanneke's result with the realizable disagreement coefficient from D., Simon and Szörenyi yields the following improved theorem, where we can replace θ_H by θ_R:

Theorem 7. *Let $\mathcal{C}$ be a concept class of VC-dimension d, D a distribution over X, $S \sim D^m$ a random sample of size m and $t \in \mathcal{C}$. Then for any $\delta > 0$ it holds with probability $\geq 1 - \delta$ for all $h \in V(S_t)$*

$$err_{t,D}(h) \leq \frac{24}{m}\left(d \log(880\, \theta_R(\mathcal{C}|t, D)) + \log \frac{12}{\delta}\right) \ .$$

Thus the sample complexity is upper bounded by $O\left(\frac{1}{\varepsilon}(d \log \theta_R(\mathcal{C}) + \log \frac{1}{\delta})\right)$.

Proof. By carefully reading the proof of Theorem 6 (as Theorem 2.23 in [9, page 51f]), one can see that $\theta_H(\mathcal{C}|t, D)$ may safely be replaced by $\theta_R(\mathcal{C}|t, D)$. Since the proof of Theorem 7 is therefore the same as Hanneke's proof for Theorem 6—except for this minor change—we will only provide a rough proof sketch for the sake of completeness:

We want to prove the following statement by induction on m: "For all $\delta > 0$ it holds with a probability of at least $1 - \delta$ over the draw of S that

$$\sup_{h \in V(S_t)} err_{t,D}(h) \leq \frac{24}{m}\left(d \log(880\, \theta_R(\mathcal{C}|t, D)) + \log \frac{12}{\delta}\right) \ ."$$

For $m \leq d$ the statement is obvious, since the error is upper bounded by 1 (and $\theta_R(\mathcal{C}|t, D) \geq 1$ always). For the inductive step from $m/2$ to m we denote by V_m and $V_{m/2}$ the version spaces after drawing m or $m/2$ points of the sample, respectively. Note that $V_m \subseteq V_{m/2}$ and therefore $\sup_{h \in V_m} err_{t,D}(h) \leq \sup_{h \in V_{m/2}} err_{t,D}(h)$.

Let $DIS := DIS(V_{m/2})$. If $D(DIS) \leq \frac{8}{m} \cdot \log \frac{3}{\delta}$, we are finished. In the other case a Chernoff bound shows that at least $n := m/4 \cdot D(DIS)$ sample points from the second $m/2$ draws lie in DIS with a probability of at least $1 - \delta/3$. We can regard these n points as a random sample drawn according to D conditioned on DIS and apply a standard PAC bound from Blumer et al. [3], showing that $\sup_{h \in V_m} err_{t,D}(h)$ is upper bounded by $D(DIS) \cdot \frac{4}{n}\left(d \log \frac{2en}{d} + \log \frac{12}{\delta}\right)$ with a probability of at least $1 - \delta/3$. To handle the occurrence of n inside the logarithm, we note that $n \leq m/4 \cdot \theta_R(\mathcal{C}|t, D) \cdot \sup_{h \in V_{m/2}} err_{t,D}(h)$ by Definition 2. After using the inductive assumption (with the probability of success set to $1 - \delta/3$) and some simplifying we arrive at the desired statement. □

We can now give our main result:

Theorem 8. *Let $\mathcal{C}$ be an intersection-closed concept class of VC-dimension d. Then the sample complexity of $\mathcal{C}$ is upper bounded by*

$$O\left(\frac{d+\log(1/\delta)}{\varepsilon}\right) \ ,$$

which matches the general lower bound and is therefore optimal. The closure algorithm realizes this bound.

Proof. The idea of the proof is similar to that of the inductive step from the proof of Theorem 6 and 7.

Let $t \in \mathcal{C}$ and D be a distribution over X and let $S \sim D^m$ be a random sample of size $m > 0$. By $h := \cap_{c \in V(S_t)} c$ we denote the closure algorithm's hypothesis after seeing S_t. We will prove that $err_{t,D}(h) \leq \frac{48}{m}(d \log 880 + \log \frac{24}{\delta})$ holds with a probability of at least $1-\delta$ over the draw of S.

Partition X into $X^+ := t$ and $X^- := X \setminus t$ and let D^+ and D^- denote D conditioned on X^+ and X^-.[5] Since h only errs on positively labeled points it holds that

$$err_{t,D}(h) = err_{t,D^+}(h) \cdot D(X^+) + \underbrace{err_{t,D^-}(h)}_{=0} \cdot D(X^-) \ . \tag{1}$$

Thus, if $D(X^+) \leq \frac{48}{m}(d \log 880 + \log \frac{24}{\delta})$ we are already finished. So assume the opposite.

Let m^+ denote the number of elements in S that lie in X^+. Obviously, m^+ is binomially distributed with an expected value of $m \cdot D(X^+)$. By our assumption $D(X^+)$ is larger than $\frac{48}{m}(d \log 880 + \log \frac{24}{\delta}) \geq \frac{8}{m} \cdot \log \frac{2}{\delta}$. From a Chernoff bound follows that

$$m^+ \geq \frac{m}{2} \cdot D(X^+) \tag{2}$$

holds with a probability of at least $1 - \delta/2$.

By Theorem 7 it holds with a probability of at least $1 - \delta/2$ that

$$err_{t,D^+}(h) \leq \frac{24}{m^+}\left(d \log(880\, \theta_R(\mathcal{C}|t, D^+)) + \log \frac{24}{\delta}\right) \ . \tag{3}$$

Since h errs on the whole disagreement region under D^+ for all S, Lemma 4 yields $\theta_R(\mathcal{C}|t, D^+) = 1$.

We now obtain the desired result by plugging (3) and (2) in (1) and observing that the overall probability of failure is upper bounded by $\delta/2 + \delta/2 = \delta$. $\square$

[5] That is, for any $S \subseteq X$, $D^+(S) := D(S \cap X^+)/D(X^+)$ and $D^-(S) := D(S \cap X^-)/D(X^-)$. The cases $D(X^+) = 0$ and $D(X^-) = 0$ are trivial: in the former case $err_{t,D}(h)$ is zero and in the latter case we can skip a part of the proof and directly apply Theorem 7 and Lemma 4.

5. Conclusions and final remarks

We improved on Hanneke's PAC bound for consistent learners and could show that the closure algorithm learns intersection-closed classes with a sample size that matches the general lower bound. To our knowledge this is the first proof of sharp sample complexity bounds for a natural family of classes in the realizable PAC-learning framework.

We demonstrated that the disagreement coefficient can be used to prove non-trivial, long-sought theorems in realizable PAC-learning. We believe that the disagreement coefficient will turn out to be helpful for other questions in computational learning theory and hope that this letter gives motivation to consider its application to related problems.

References

[1] L. G. Valiant, A theory of the learnable, in: STOC, 1984, pp. 436–445.

[2] A. Ehrenfeucht, D. Haussler, M. Kearns, L. Valiant, A general lower bound on the number of examples needed for learning, in: COLT, 1988, pp. 139–154.

[3] A. Blumer, A. Ehrenfeucht, D. Haussler, M. K. Warmuth, Learnability and the Vapnik-Chervonenkis dimension, Journal of the ACM 36 (4) (1989) 929–965.

[4] M. K. Warmuth, The Optimal PAC Algorithm, in: COLT, 2004, pp. 641–642.

[5] P. Auer, R. Ortner, A new PAC bound for intersection-closed concept classes, Machine Learning 66 (2007) 151–163.

[6] E. Giné, V. Koltchinskii, Concentration inequalities and asymptotic results for ratio type empirical processes., The Annals of Probability 34 (3) (2006) 1143–1216.

[7] K. S. Alexander, Rates of growth and sample moduli for weighted empirical processes indexed by sets, Probability Theory and Related Fields 75 (3) (1987) 379–423.

[8] S. Hanneke, A bound on the label complexity of agnostic active learning, in: ICML, 2007, pp. 353–360.

[9] S. Hanneke, Theoretical Foundations of Active Learning, Ph.D. thesis, Carnegie Mellon University (2009).

[10] M. Darnstädt, H. U. Simon, B. Szörényi, Supervised Learning and Co-training, Theoretical Computer Science 519 (2014) 68–87.

[11] V. N. Vapnik, A. Y. Chervonenkis, On the Uniform Convergence of Relative Frequencies of Events to Their Probabilities, Theory of Probability and its Applications XVI (2) (1971) 264–280.